252

THE RICE INSTITUTE PAMPHLET

Vol. XXVII January, 1940 No. 1

Published by
THE RICE INSTITUTE

A university of liberal and technical learning
founded by William Marsh Rice in the City
of Houston, Texas, and dedicated by him to
the advancement of Letters, Science, and Art

The Rice Institute Pamphlet is published at Houston, Texas, and is issued in January, April, July, and October. It was entered as second-class matter, April 15, 1915, at the post-office at Houston, Texas, under the act of August 24, 1912

A3
36
W45
Vol.27

THE
RICE INSTITUTE PAMPHLET

Vol. XXVII January, 1940 No. 1

Rice University studies,

30 Aug 68

Published by
THE RICE INSTITUTE

A university of liberal and technical learning
ounded by William Marsh Rice in the City
f Houston, Texas, and dedicated by him to
he advancement of Letters, Science, and Art

1127-25

THE RICE INSTITUTE PAMPHLET

Vol. XXVII January, 1940 No. 1

CONTENTS

PAGE

Three Lectures on Mathematical Subjects—Delivered at the Rice Institute, December 11, 12, and 13, 1939, by Karl Menger, Ph.D. (Vienna), Professor of Mathematics at the University of Notre Dame.

I. Analysis and Metric Geometry 1
II. On Algebra of Geometry and Recent Progress in Non-Euclidean Geometry 41
III. Topology Without Points 80

I

ANALYSIS AND METRIC GEOMETRY

LINE INTEGRALS, THEIR SEMICONTINUITY PROPERTIES, AND THEIR INDEPENDENCE OF THE PATH

1. THE LENGTH (RIEMANN INTEGRATION, LEBESGUE INTEGRATION, WEIERSTRASS-ARCHIMEDES SUMS)

ONE of the oldest and best known line integrals is the length of a curve. According to the classical theory, the curve C in the (t, y)-plane given by the equation $y = y(t)$, $(a \leq t \leq b)$ has the length

$$l(C) = \int_a^b \sqrt{1 + (y'(t))^2}\, dt.$$

According to Riemann's definition of a definite integral, $l(C)$ has the following meaning: Let T be a finite ordered subset of $[a, b]$, say

$$T = \{t_0, t_1, \cdots t_{n-1}, t_n\},$$

where $t_0 = a$, $t_n = b$, and $t_i < t_{i+1}$. We call the greatest of the n numbers $t_{i+1} - t_i$ the norm of T and denote it by $\nu(T)$. Let T^* be a set $\{t_0, t_0^*, t_1, t_1^*, \cdots t_{n-1}, t_{n-1}^*, t_n\}$ where $t_i \leq t_i^* \leq t_{i+1}$. By the *Riemann sum* associated with T^*, we mean the sum

$$R(T^*) = \sum_{i=0}^{n-1} \sqrt{1 + (y'(t_i^*))^2}\,(t_{i+1} - t_i).$$

The length $l(C)$ is that number (finite or $+\infty$) to which the numbers $R(T^*)$ come as close as we please for all T^* for which $\nu(T)$ is sufficiently small—provided that a number of this kind exists. It does exist for the curve $y=y(t)$ if and only if the integrand $\sqrt{1+(y'(t))^2}$ or, what is equivalent, the function $y'(t)$ is continuous almost everywhere (i.e., for all t outside of a set of measure 0 in Lebesgue's sense).

In particular, the integrand and hence the derivative $y'(t)$ must exist everywhere in $[a, b]$; that is to say, the curve C must have a tangent at each point. By a well-known procedure of splitting the domain of integration into parts, we may admit a finite set of numbers t for which $y'(t)$ is not defined, and even certain simple infinite sets of points at which the curve C has no tangents.

At any rate, the classical theory ties up the definition of the length of a curve with the question of whether the curve has tangents and how these tangents vary.

This is unsatisfactory, because the primitive geometric ways of determining the length of a curve do not make use of tangents to the curve. Our intuitive idea of the length has no connection with that of tangents. Moreover, we are interested in the length of any curve regardless of whether or not it has tangents, and we actually can determine the length of curves which do not have tangents. There are two ways of severing the ties between length and tangents.

One attempt consists in interpreting the $\int$ sign in $l(C) = \int_a^b \sqrt{1+(y'(t))^2}dt$ as a Lebesgue integral. The Lebesgue integral of a function in an interval may exist without the function being defined everywhere in the interval. It is sufficient that the integrand is defined almost everywhere provided that, on its domain of definition, it satisfies certain conditions. Now if a curve C given by $y=y(t)$ $(a \leq t \leq b)$ has a finite length $l(C)$ in the sense of geometry, then Lebesgue proved that the set of all t for which C has no tangent is of measure 0. Thus, for such a curve, $y'(t)$ and $\sqrt{1+(y'(t))^2}$ do exist almost everywhere. Moreover, Lebesgue proved that if $y(t)$ defines a curve of finite length then the function $\sqrt{1+(y'(t))^2}$ has a finite Lebesgue integral. But with regard to the relation between this integral and the length $l(C)$, all that can be said is that the integral is never greater than the length. It may be actually smaller. Tonelli proved that the equality $l(C)=\int_a^b\sqrt{1+(y'(t))^2}dt$ holds if and only if the function $y(t)$ is absolutely continuous.[1]

For instance, let C be the curve joining the points $(0, 0)$ and $(1, 1)$ given by the equation $y=y(t)$ $(0\leq t\leq 1)$, where $y(t)$ is the continuous but not absolutely continuous function, assuming the value $1/2$ for $1/3\leq t\leq 2/3$, the values $1/4$ for $1/9\leq t\leq 2/9$ and $3/4$ for $7/9\leq t\leq 8/9$, etc., in general the value

$$y=\sum_{k=1}^{n-1}\frac{a_k}{2^k}+\frac{1}{2^n} \quad \text{for} \quad \sum_{k=1}^{n}\frac{2a_k}{3^k}+\frac{1}{3^n}\leq t\leq \sum_{k=1}^{n}\frac{2a_k}{3^k}+\frac{2}{3^n}$$

$$(a_1,\ a_2,\ \cdots\ a_{n-1}=0 \text{ or } 1).$$

This function $y(t)$ being constant on intervals whose total length is 1 has the derivative 0 almost everywhere in $[0, 1]$. Thus the Lebesgue integral

$$\int_0^1\sqrt{1+(y'(t))^2}dt=\int_0^1 1\cdot dt=1.$$

This number is even smaller than the length $\sqrt{2}$ of the straight segment between the endpoints $(0, 0)$ and $(1, 1)$ of the curve C. The length of C is 2.

Thus, by interpreting $\int$ as a Lebesgue integral, we cannot uphold the definition of the length of a curve as the line integral $\int\sqrt{1+(y'(t))^2}dt$ beyond a special class of curves of finite length, viz., the curves defined by absolutely continuous functions—although the line integral itself can be formed for more general curves. Moreover, this method digresses from the geometric intuition that accompanies our elementary determination of length.

The second way out of the difficulty consists in considering what is called Weierstrass sums instead of Riemann sums. By the *Weierstrass sum* associated with T, we mean an expression similar to a Riemann sum, but involving difference quotients instead of differential quotients, namely

$$W(T)=\sum_{i=0}^{n-1}\sqrt{1+\left(\frac{y(t_{i+1})-y(t_i)}{t_{i+1}-t_i}\right)^2}\;(t_{i+1}-t_i).$$

If for $y(t)$ the mean value theorem of differential calculus holds, then for each T there exists a T^* such that $R(T^*)=W(T)$. Hence, if there is a number $l(C)$ to which $R(T^*)$ converges when $\nu(T)$ converges toward zero, then this number $l(C)$ is also the limit of $W(T)$ when $\nu(T)$ converges toward zero. But there may exist a number $l(C)$ to which $W(T)$ converges with $\nu(T)\to 0$ without $R(T^*)$ approaching a limit. In fact, one can prove for each curve C which is given by a continuous function $y=y(t)$, that there is a number $l(C)$ finite or $+\infty$ to which $W(T)$ converges with $\nu(T)\to 0$. However, as we saw, for many continuous functions $y(t)$, there is no number to which $R(T^*)$ would converge with $\nu(T)\to 0$. E.g., there is no such number $l(C)$ if the continuous function $y(x)$ is nowhere differentiable, for the simple reason that in this case the numbers $R(T^*)$ themselves do not exist.

The second way of severing the definition of the length of a curve from its differentiability properties consists in calling length of C the limit of the numbers $W(T)$ rather than that of the numbers $R(T^*)$.

If we examine $W(T)$, we see that it is equal to

$$\sum_{i=0}^{n-1}\sqrt{(t_{i+1}-t_i)^2+(y(t_{i+1})-y(t_i))^2}.$$

In the Euclidean plane the i-th summand of this sum is the distance between p_i and p_{i+1}, where we denote by p_i the point with the coordinates $(t_i, y(t_i))$. Thus the sum $W(T)$ is the length of the polygon $\{p_0, p_1, \cdots p_n\}$ inscribed in the curve C. It is what may be called *Archimedes Sum*

$$A(T) = \sum_{i=0}^{n-1} \text{dist}\,(p_i, p_{i+1}),$$

for it is these lengths of inscribed polygons that, in special cases, Archimedes used for the determination of the length of a curve C.

Now when we spoke about primitive geometric methods of determining the length of a curve, we meant the limit to which the lengths $A(T)$ of the inscribed polygons converge when the norm of the polygons tends toward 0. And when the theory of length as a Lebesgue line integral states that $\int_a^b \sqrt{1+(y'(t))^2}dt$ exists for each curve of finite length, and admits that this integral may be smaller than the *geometric* length of the curve, it is again this limit of the $A(T)$ which is meant. Thus the identity of $W(T)$ with $A(T)$ seems to indicate that the Weierstrass sums are a tool of analysis more adequate to the geometric problem than the Riemann or Lebesgue sums. At the same time, a theory based on the $W(T)$ is the most radical way of severing the concept of length from that of tangents, and of freeing the analysis from unessential differentiability assumptions.

2. THE GENERAL LINE INTEGRAL AND FOUR MAIN PROBLEMS CONCERNING IT

Now we consider the general line integral. In the following we shall restrict ourselves to curves in the Euclidean plane since the generalization to curves of higher dimensional Euclidean spaces does not present any difficulty. In the plane we shall admit any curve C given by a parametric representation $x=x(t)$, $y=y(t)$ $(a \leq t \leq b)$. A line integral is determined by a function of four real variables $F(x,y,u,v)$. The value of the line integral of F along the curve C is

$$J(C)=\int_a^b F(x(t),\ y(t),\ x'(t),\ y'(t))dt.$$

Some of the main problems concerning these general line integrals are the following questions:

I. Under what conditions concerning the integrand F and the curve C does $J(C)$ *exist*?

II. Under what conditions concerning F is the integral a *semicontinuous* functional of C, say lower semicontinuous, i.e., a functional with the property that for each curve C and each $\epsilon>0$ there exists a neighborhood of C such that for each curve C_1 located in this neighborhood we have $J(C_1)>J(C)-\epsilon$?

III. Under what conditions concerning a class of curves and the integrand F does the class contain a curve *minimizing* $J(C)$, i.e., a curve for which the value of J does not surpass the value of J for any other curve of the class?

IV. Under what conditions concerning the integrand F is the integral $J(C)$ *independent of the path*, that is to say, assumes the same value for any two coterminal curves?

3. THE ANSWER TO THE FUNDAMENTAL QUESTIONS ON THE BASIS OF RIEMANN INTEGRATION

The classical treatment of these problems is based on Riemann integration. From this point of view $J(C)$ is the limit of Riemann sums

$$R(T^*)=\sum_{i=0}^{n-1} F(x(t_i^*),\ y(t_i^*),\ x'(t_i^*),\ y'(t_i^*))(t_{i+1}-t_i)$$

when $\nu(T)\rightarrow 0$. In order that such a limit exists, it is necessary and sufficient that the function $F(x(t), y(t), x'(t), y'(t))$ is almost everywhere continuous. To guarantee this condition, we must not only know that F has certain continuity properties, but also that $x'(t)$ and $y'(t)$ exist and are continuous to a large extent. In answer to question I, the classical theory did not hesitate to make these assumptions, but, of course, this procedure is no more satisfactory in the general case, than it is in the case of length.

It is clear that similar restrictions would have to be imposed on F and C in order to guarantee that the Riemann integral is a semicontinuous functional and to prove the existence of minimizing curves in answer to questions II and III, although these questions seem not to have been studied extensively for Riemann integrals.

To question IV the classical theory gave two answers. Let $u(x, y)$ and $v(x, y)$ be two continuous functions. Then $\int udx+vdy$ is independent of the path if either one of the two following conditions is satisfied:

1. The partial derivatives $\frac{\partial u}{\partial y}$ and $\frac{\partial v}{\partial x}$ exist, are continuous, and are equal to each other at each point (x, y) of a closed domain.

2. $\frac{\partial u}{\partial y}=\frac{\partial v}{\partial x}$, and both functions u and v have differentials at each point. A function $g(x, y)$ is said to have a differential at the point (x_0, y_0), if there exist two numbers $a(x_0, y_0)$ and $b(x_0, y_0)$ with the following property: for each $\epsilon>0$ there is a $\eta>0$ such that

$$\begin{aligned} g(x, y) = g(x_0, y_0) &+ a(x_0, y_0)(x-x_0) \\ &+ b(x_0, y_0)(y-y_0)+\zeta(x,y)(|x-x_0|+|y-y_0|) \\ \text{and } |\zeta(x, y)| &< \epsilon \end{aligned}$$

for each (x, y) for which $|x-x_0|+|y-y_0|<\eta$. The numbers $a(x_0, y_0)$ and $b(x_0, y_0)$ are, of course, the partial derivatives of g for (x_0, y_0) with respect to x and y, respectively.

The sufficiency of condition 2 was proved by Heffter, following Goursat's method in the theory of the functions of a complex variable. This condition has a local character since the dependence of η on ϵ may vary with (x_0, y_0). No uniformity of the extent to which $g(x, y)-g(x_0, y_0)$ is approximated with a prescribed degree of accuracy by the linear terms $a(x_0, y_0)(x-x_0)+b(x_0, y_0)(y-y_0)$, is postulated. The older condition 1, by assuming continuity of the partial derivatives in a bounded and closed domain, does imply uniform continuity of these partial derivatives. Exceptional points at which the differentials or the partial derivatives do not exist or the partial derivatives are discontinuous and unequal are admissible on the basis of Riemann integration, provided that the set of all exceptional points has the content 0, i.e., may be covered by a finite number of squares of an arbitrarily small total area.[2]

Summarizing, we can say: with regard to all four problems, the classical theory based on Riemann integration made many assumptions about the differentiability of the integrand F and the curve C which are required by the method rather than by the questions themselves. For, the questions concerning the line integrals can be formulated with regard to integrands that are not differentiable and curves that have no tangents.

4. THE ANSWER TO THE FOUR FUNDAMENTAL QUESTIONS ON THE BASIS OF LEBESGUE INTEGRATION

In the theory of length we saw two ways out of this difficulty. With regard to general line integrals, the first of these ways was followed. One considered Lebesgue line integrals.

Hahn seems to have been the first to introduce Lebesgue integrals into the calculus of variations.[3] Tonelli based his entire book[1] on Lebesgue integration. The fundamental concepts underlying this work which marked one of the greatest steps forward in its field can be summarized as follows:

a. *The Integrand.* Tonelli starts by making Weierstrass's assumption that the integrand $F(x, y, u, v)$ is positively homogeneous with regard to u and v, i.e., satisfies the equality $F(x, y, ku, kv) = kF(x, y, u, v)$ for each $k > 0$ and each quadruple x, y, u, v for which (x, y) lies in a certain domain of the plane. This assumption is made in order to guarantee that for any two different parametric representations of the same curve, the integral assumes the same value. It implies that

$$\frac{F_{uu}}{v^2} = -\frac{F_{uv}}{uv} = \frac{F_{vv}}{u^2}$$

where we write F_{uu} for $\frac{\partial^2 F}{\partial u^2}$, etc. Calling $F_1(x, y, u, v)$ the common value of the above three quotients Tonelli assumes that $F_1 \geq 0$ or, as he says, that F is *positively quasi-regular.* Furthermore, he assumes that the function $F(x, y, u, v)$ is *continuous* with respect to the quadruples of variables. Moreover, F is assumed to admit at least *first partial derivatives* with respect to the last two variables. For part of his theory Tonelli assumes the existence of second derivatives, but remarks that other parts are independent of this assumption. For instance, the hypothesis of quasi-regularity which involves second partial derivatives can be replaced by the assumption that for each point (x_0, y_0) the figurative of F is convex downward. The figurative of F for the point (x_0, y_0), introduced by Hadamard, is the surface $z = F(x_0, y_0, u, v)$ in the (u, v, z)-space. On account of the positive homogeneity of F, this surface is a cone with the vertex at the origin. The cone degenerates into a plane if and only if $F_1(x_0, y_0, u, v) = 0$ for all values of u and v.

A subsequent idea due to McShane[4] and Aronszajn[5] allows us to dispense even with continuity of the integrand F provided that F is lower semicontinuous, and the figurative convex at each point.

b. *The Admissible Curves.* The curves along which the Lebesgue line integral $J(C)$ may be formed are the curves of finite length, if this length is used as parameter. That is to say, we can form $\int_0^{l(C)} F(x(s), y(s), x'(s), y'(s))ds$ for each curve C given by two equations $x = x(s)$, $y = y(s)$ $(0 \leq s \leq l(C))$ where s denotes the length of the segment of the curve C between the initial point $(x(0), y(0))$ and the point $(x(s), y(s))$. If C is given by two equations $x = x(t)$, $y = y(t)$ $(a \leq t \leq b)$, then Tonelli requires absolute continuity of the functions $x(t)$, $y(t)$, a condition which is automatically satisfied by the functions $x(s)$, $y(s)$ defining a curve of finite length.

Under the specified conditions concerning the integrand F, Tonelli proves that $J(C)$, the integral of F, is lower semicontinuous for all curves C of uniformly bounded lengths. Under slightly stronger conditions concerning F, he proves that the integral is lower semicontinuous for all curves of finite length. The additional assumption concerning F (called *seminormality* of F) is that for no point (x_0, y_0) is $F_1(x_0, y_0, u, v) = 0$ for all (u, v) or, in other words, that for no point is the figurative of F a plane. Tonelli even admits points (x_0, y_0) for which the figurative is a plane provided that each of these points lies within a neighborhood such that $J(C) \geq 0$ for any closed curve C of finite length contained in the neighborhood.

c. *The Functional of Comparison.* As we saw when speaking about the admissible curves, the integral along a curve is frequently related to the length of the curve. The integral is merely defined for curves of finite length; some semicontinuity properties are merely proved on the sets of curves of uniformly bounded lengths; and, as we shall see, Tonelli's theorem about the existence of curves that minimize the integral, contains another reference to the length. For these reasons we shall call the length of curves the functional of comparison used in Tonelli's theory.

d. *The Space.* All the curves considered in Tonelli's theory are contained in a Euclidean or at least a Cartesian space in which each point can be described by n real coordinates.

In answer to question III Tonelli states conditions sufficient for the existence in each closed class of curves of a curve C_0 that minimizes the integral $J(C)$. From the lower semicontinuity of the integral he derives the existence of such a minimizing curve under the additional hypothesis that for each finite number M the curves C for which $J(C) \leq M$ are of uniformly bounded lengths. Implicitly underlying Tonelli's deduction is a theorem of the general analysis generalizing the classical theorem of R. Baire that a lower semicontinuous function on a compact domain of definition assumes its minimum.

A construction due to Hahn and later generalized by Carathéodory, Tonelli, Graves and McShane throws light on the additional hypothesis that for each M the curves C for which $J(C) \leq M$, are of uniformly bounded lengths. If this condition is not satisfied (i.e., in case there exists a sequence of curves $C_1, C_2, \cdots$ for which all numbers $J(C_n)$ are $\leq M$ while the lengths of $C_1, C_2, \cdots$ converge toward ∞), then Hahn's construction yields a sequence of curves $D_1, D_2, \cdots$ for which all the numbers $J(D_n)$ are ≤ 0 while the lengths of $D_1, D_2, \cdots$ converge toward ∞. Hence it is sufficient to postulate that the curves C for which $J(C) \leq 0$ are of uniformly bounded lengths.

With regard to problem IV, Montel introduced Lebesgue integration and proved that a line integral $\int udx+vdy$ is independent of the path provided that $\partial u/\partial y$ and $\partial v/\partial x$ exist and are equal almost everywhere.[6] Montel's successors weakened his assumptions further, in particular with regard to functions of a complex variable, the conditions being formulated in terms of the partial derivatives of u and v.

Of a different type is a condition due to Schauder.[7] Besides continuity of the functions $u(x, y)$ and $v(x, y)$ defined in a rectangle $a \leq x \leq b$, $c \leq y \leq d$, it requires that

$$\int_c^y v(x, \eta)d\eta - \int_a^x u(\xi, y)d\xi,$$

a function of the variables x and y, can be represented as the sum of two functions $\varphi(x)$ and $\psi(y)$ each depending upon one of the two variables only. This very elegant analytic condition does not presuppose differentiability of u and v. On the other hand, it does not seem to give us much of a geometric insight into where the classical arguments are redundant. The work of Besikovic, Looman, Menchoff, Saks and others is based on differentiability assumptions.

In terminating this brief summary of how the four main questions were treated for Lebesgue line integrals, it should be emphasized that with Lebesgue integration, a branch of modern geometry was introduced into Analysis, viz., the geometry of measure. What fundamental role this geometry played in the modern analysis one will appreciate by reading, for instance, Saks's book on integration.[1]

At any rate, the progress marked by the introduction of

Lebesgue integration as compared with the old theory based on Riemann integrals was so great that it can hardly be overestimated. Today, in view of its permanent achievements, the theory developed on this basis during the first two decades of this century can be called classical.

5. EXAMPLES OF CASES EXCLUDED BY THE CLASSICAL THEORIES

It seems to me, however, that there are two objections to this method. The first is that although the method is based on a chapter of geometry, viz., the theory of measure, it is not capable of as simple and intuitive geometric interpretation as, e.g., the theory of length based on the study of inscribed polygons. Facts like the difference between geometric length and the length integral of Lebesgue discussed in section 1 of this paper are merely symptoms of a more general deviation of analysis from geometric intuition in this particular field. But since this argument is one of personal taste rather than one of objective criticism, I shall not insist on it.

The second remark is of a quite objective character. Although by introducing the Lebesgue integral one weakens considerably the assumptions that the theory dealing with Riemann integrals had to make concerning the integrand F and the curve C, one still has to introduce assumptions that have no direct connection with the problem and are introduced for the sake of the method rather than for the sake of the question. This is not only an esthetic disadvantage, but has very concrete mathematical consequences that are undesirable.

Examples to illustrate this remark will be clearer if compared with one of the oldest problems of the calculus of variations as a background. We mean a question that does not present any difficulties on the basis of Riemann or Lebesgue integration, but would do so for more special kinds of integrals, viz., the problem of finding the curve $x(t)$, $y(t)$, $(a \leq t \leq b)$, which joins two given points $p_0 = (x_0, y_0)$ and $p_1 = (x_1, y_1)$ and which, rotated around the x-axis, generates a surface of revolution of minimum area. Analytically speaking, we are looking for a pair of functions $x(t)$, $y(t)$, $(a \leq t \leq b)$, satisfying the conditions $x(a) = x_0$, $y(a) = y_0$, $x(b) = x_1$, $y(b) = y_1$ and minimizing the integral

$$\int_a^b y(t)\sqrt{[x'(t)]^2 + [y'(t)]^2}\,dt.$$

If the points p_0 and p_1 are sufficiently far apart, and sufficiently close to the x-axis, then the minimizing curve discovered by Goldschmidt is a broken line with two corners on the x-axis. This case is, of course, taken care of by introducing Lebesgue integration and it can be handled even on the basis of Riemann integration since the set of points at which the integrand is not defined consists of two points only. But if at the start we restricted ourselves to the consideration of functions $x(t)$ and $y(t)$ that are everywhere differentiable or, geometrically speaking, of curves with a tangent at each point, then we should exclude just that curve in which we are most interested, viz., the one that minimizes the integral. If we based the theory on a concept of integration applicable only to functions which are everywhere differentiable, the effect would be the same.

As a first example to illustrate our criticism, we consider the problem of finding two functions $x(t)$, $y(t)$ $(a \leq t \leq b)$ satisfying the conditions $x(a) = 0$, $y(a) = 0$, $x(b) = 1$, $y(b) = 0$ and minimizing the integral

$$\int_a^b \sqrt{(x^2+y^2)^3(x'^2+y'^2) + 2(x^2+y^2)(xy'-yx')(xx'+yy') + (xx'+yy')^2}\,dt$$

For any curve of finite length joining the point (0, 0) with the point (1, 0) the integral can be shown to assume a positive value. However, for some spirals of infinite length joining the two points the (improper) integral assumes the value zero. The existence of problems in which the minimizing curves are of infinite length was discovered by Hahn. The simple integrand mentioned above, a square root of a polynomial of degree eight, is due to Carathéodory.[8] The examples of this type show that by restricting ourselves to the consideration of curves of finite lengths at the start, we sometimes exclude just those curves in which we are mainly interested, viz., the curves minimizing the integral. On the other hand, the classical theory, whether based on Riemann or on Lebesgue integration, had to restrict itself to curves of finite lengths, or still more special curves. A method enabling us to define line integrals along all continuous curves, and to prove semicontinuity properties in such an extended domain would thus mark a distinct progress.

As a second example[9] we remember that in studying the question when $\int udx+vdy$ is independent of the path the classical theory restricts itself to functions u and v which admit partial derivatives, at least almost everywhere. Only in this case the classical condition of Cauchy, $\frac{\partial u}{\partial y}=\frac{\partial v}{\partial x}$, can be formulated. Now let $f(z)$ be a continuous function of one real variable z and set $u(x, y)=f(x+y)$ and $v(x, y)=f(x+y)$. Then the $\int udx+vdy$ is equal to $\int f(x+y)(dx+dy)=\int f(x+y)\, d(x+y)$. This integral is certainly independent of the path. For if $F(z)$ is a function satisfying the condition $F'(z)=f(z)$, then the value of the integral along a curve is equal to the difference of the values of F at the two end points of the curve. Now assume that the continuous function $f(z)$ is nowhere differentiable. Then the

$\int u\,dx+v\,dy$ will be independent of the path; although neither u nor v has any partial derivative.

6. WEIERSTRASS SUMS AND GEOMETRY OF DISTANCE

These and other examples suggested another way out of the difficulties, a way parallel to the second one mentioned in the special case of the length. We shall consider the integral neither in Riemann's nor in Lebesgue's sense, but as the limit of Weierstrass sums. By a Weierstrass sum of the integral $\int_a^b F(x(t), y(t), x'(t), y'(t))dt$ associated with the set T we mean

Weierstrass sum $W(T)=$

$$\sum_{i=0}^{n-1} F\left(x(t_i), y(t_i), \frac{x(t_{i+1})-x(t_i)}{t_{i+1}-t_i}, \frac{y(t_{i+1})-y(t_i)}{t_{i+1}-t_i}\right)(t_{i+1}-t_i).$$

Weierstrass considered the integral as the limit to which these sums converge when $\nu(T)\to 0$. Osgood continued the work for the sake of eliminating superfluous differentiability conditions concerning the admissible curves. Bolza devotes a paragraph of his book to this idea. Tonelli studied the Weierstrass sums in a paper. But no one seems to have been interested in the general conditions concerning both the integrand and the admissible curves, to say nothing of the functional of comparison and the space, under which such a limit exists and is semicontinuous. The subsequent development of the calculus of variations moved almost entirely along the lines of Lebesgue integration. Weierstrass sums were not thoroughly studied until a few years ago, when Bouligand[10] suggested a definition of the integral as a limit of these sums, and independently the writer of this paper in a series of publications[11] developed an extensive theory of these limits applying methods developed in his former papers on metric geometry, or geometry of distance.[12]

The connection between these questions and the geometry of distance is established by a step corresponding to that from Weierstrass sums to Archimedes sums in the case of the length. Denoting by p_i the point with the coordinates $x_i = x(t_i)$, $y_i = y(t_i)$, and by $\theta_{i,\,i+1}$ the angle which the vector from p_i to p_{i+1} includes with the positive x-axis, we can write the Weierstrass sum for the integrand F associated with the finite set T in the following way:

$$\sum_{i=1}^{n-1} F(x_i, y_i, \cos\theta_{i,\,i+1}, \sin\theta_{i,\,i+1}) \times \text{Eucl. dist. } p_i p_{i+1}$$

where × is a multiplication sign, and Eucl. dist. $p_i p_{i+1}$ denotes the Euclidean distance between p_i and p_{i+1}.

The expression can be further transformed if for any two points $p = (x, y)$ and $q = (x', y')$ we briefly write $F(p, \theta_{pq})$ instead of $F(x, y, \cos\theta_{pq}, \sin\theta_{pq})$ where θ_{pq} denotes the angle between the vector from p to q and the positive x-axis, and furthermore set $F(p, \theta_{pq}) = 0$, if $p = q$. In this notation we associate with any two points p and q of the Euclidean plane a new distance which we shall call the F-distance between p and q, viz., the number

$$F\text{-dist. } pq = F(p, \theta_{pq}) \times \text{Eucl. dist. } pq.$$

If an individual problem of the calculus of variations determined by an integrand $F(p, \theta)$ is under consideration, then we also shall call the F-distance the *variational distance.* In this notation the Weierstrass sum can be written in the following form entirely analogous to an Archimedes sum

$$\sum_{i=1}^{n-1} F\text{-dist. } p_i p_{i+1} \quad \text{or} \quad \sum_{i=1}^{n-1} \text{var. dist. } p_i p_{i+1}.$$

In other words, each Weierstrass sum is the length of a polygon inscribed into the curve C if the determination of the length is based on what we called the F-distance or the variational distance of pairs of points. The functional that the writer's theory associates with the curve C is the limit which the variational F-lengths of the polygons inscribed in C approach when the norms of the polygons approach zero—if there is such a limit. Our functional thus may be considered as the *F-length* or *variational length of the continuous curves*, for our derivation of the functional from the variational F-distance is exactly the same as the derivation of the Euclidean length of a curve from the Euclidean distance.

If $F(p, \theta)=1$ for each point p and each direction θ, then F-dist. $pq =$ Eucl. dist. pq for any two points p, q, and the variational F-length of each curve is its Euclidean length.

There is, however, one important difference between the Euclidean length and the general F-length of curves: Each curve of a Euclidean space has a finite or infinite Euclidean length, that is to say, for each curve the Euclidean lengths of the inscribed polygons converge toward a finite number or toward $+\infty$ when the norms of the polygons tend toward 0. But for certain integrands F there are curves for which the F-lengths of the inscribed polygons do not converge at all (neither toward a finite nor toward an infinite number) when the norms tend toward 0.

As an illustration of this fact, F. Alt gave the following example[13] of a continuous curve C and a continuous quasi-regular integrand F. The curve C is contained in the x-axis, and given by the equation $x(t)=t^2 \cos 2\pi/t^6$, $y(t)=0$, $(-\sqrt[3]{2} \leq t \leq 0)$. The integrand is $F(x, \theta) = -\cos\theta \cdot x^2 \cos 2\pi/x^3$. Among the polygons of arbitrarily small norms inscribed into C there are some whose F-lengths are arbitrarily small, and some whose F-lengths are arbitrarily large. Hence, C has no F-length. Owing to its greater complexity the theory of the F-length presents difficulties which do not arise in the special case of the Euclidean length.

In the following five sections we shall deal with the application of metric geometry to the three problems about line integrals which are of interest for the calculus of variations: When does the integral exist? When is it semicontinuous? When do there exist minimizing curves? We shall see that the metric methods result in generalizations of all the four basic concepts: the integrand, the admissible curves, the functional of comparison, the space. After that we shall apply our method to the question of when the integral is independent of the path.

7. METRIC METHODS AND THE INTEGRAND

Like Tonelli and his successors we shall need continuity and regularity assumptions concerning $F(p, \theta)$. But we shall not even need lower semicontinuity of F at each point: We admit that each curve of finite length may contain a set of measure 0 consisting of points p at which $F(p, \theta)$ is not even lower semicontinuous provided that each of these exceptional points is contained in a neighborhood such that $J(C) \geq 0$ for each closed curve C contained in this neighborhood. Here, following Carathéodory, a set is said to be of measure 0 if for each $\epsilon > 0$ it can be covered by a denumerable sequence of spheres for which the sum of the diameters is $< \epsilon$. A set which can be covered by a finite set of spheres for which the sum of the diameters is arbitrarily small, will be said to be of content 0.

Moreover, we shall not need quasi-regularity of F at each point, i.e., convexity of the figurative of F for each point (x_0, y_0). For concave surfaces we can define a degree of concavity and then admit that each curve of finite length for each $\tau > 0$ may contain a set of content 0 consisting of points for which the figurative of F has a degree of concavity $\geq \tau$ provided that the degree of concavity is uniformly bounded along the whole curve.[14]

8. METRIC METHODS AND THE ADMISSIBLE CURVES

Under the aforementioned assumptions about the integrand F, the variational F-length $J(C)$ exists for each curve of finite Euclidean length and is lower semicontinuous on each set of curves whose Euclidean lengths are uniformly bounded.

Under slightly stronger conditions we can prove that the functional $J(C)$ exists *for all continuous curves* whether of finite or of infinite Euclidean length, and is lower semicontinuous for all continuous curves. The additional assumption is that for each point p there exists a neighborhood of p and a number $\omega(p) > 0$ such that for each closed polygon contained in the neighborhood the variational length is at least the $\omega(p)$-fold of its Euclidean length.[15] If F is continuous, quasi-regular, and seminormal, then it can be proved that this additional assumption is satisfied. Moreover, our theory formulates geometric conditions guaranteeing that integrands which are neither continuous nor quasi-regular, satisfy the additional assumption.[16]

While these theorems enable us to define line integrals of certain integrands and to prove their semicontinuity for all continuous curves, they are complemented by a theorem which impairs their use for the problem of finding curves which minimize the integral. For it can be shown[17] that under the aforementioned conditions along each curve whose Euclidean length is infinite the integral of F has the value $+\infty$.

However, our theory enables us to form the integral along any continuous curve even for certain integrands that are not seminormal and do not satisfy the additional assumption mentioned above. And for such integrands, the integral along a curve of infinite Euclidean length may be finite.[18] The integrands of this type which our theory enables us to deal with form a fairly large class of functions including, in particular, the integrands discovered by Hahn and Carathéodory, which we discussed in Section 5.

9. METRIC METHODS AND THE FUNCTIONAL OF COMPARISON

The reason why Tonelli and his successors referred to the length as a functional of comparison was that they had to make use of the following theorem due to Hilbert: In a compact space for each finite number M the set of all continuous curves whose lengths are $\leq M$, is compact. That is to say, from each sequence of curves whose lengths are $\leq M$ we can extract a subsequence converging toward a continuous curve. (On account of the lower semicontinuity of the length this limit curve likewise has a length $\leq M$.) The following generalization of Hilbert's theorem has been proved by the writer.[19]

In a compact space let $L(C)$ be any lower semicontinuous functional, defined (although not necessarily finite) for all continuous curves C and having the following properties:

(1) If C is a curve for which $L(C) < \infty$, and p is a point of C, then we have $L(C')+L(C'')=L(C)$ where C' and C'' denote the initial and terminal segment into which C is divided by p. If p converges toward the terminal point of C, then $L(C')$ converges toward $L(C)$.

(2) For any two distinct points of the space, p and q, the greatest lower bound of the numbers $L(C)$ for all continuous curves C joining p and q is >0.

(3) Each point p, for each $\epsilon>0$, can be joined to each point of a sufficiently small neighborhood by a continuous curve C for which $L(C)<\epsilon$.

Then for each finite number M the set of all continuous curves C for which $L(C) \leq M$ is compact.

By virtue of this theorem, instead of the length we may use as a functional of comparison any lower continuous functional $L(C)$ with the three aforementioned properties. In particular, we may use many line integrals. For, if $G(p, \theta)$ is any function of a point p and a direction θ, and $L(C)$ is the line integral of G along C, then the functional $L(C)$ certainly satisfies the condition (1). If along each curve of finite length the integrand G is almost everywhere >0, then the line integral also satisfies condition (2). If, roughly speaking, almost everywhere along each curve of finite length, G is lower semicontinuous and has a figurative that is convex upward, then, as we know, the line integral of G is a lower semicontinuous functional. We have only to postulate that condition (3) holds, which, in fact, is a condition concerning the space (the domain of definition) rather than the functional. It states a property of the space similar to the local connectedness studied in topology. If $L(C)$ denotes the diameter of C, then condition (3) postulates that each point can be joined to each sufficiently close neighbor point by a continuous curve of arbitrarily small diameter, and that is exactly the property described in topology by the words "the space is locally connected at each point." Hence, we shall express the condition (3) by saying that the space is L-locally connected.

In particular, we can say: If G is a function which along each curve of finite length is almost everywhere >0, lower semicontinuous and quasi-regular, and the space is L-locally connected with regard to the integral $L(C)$ of G along C, then the functional $L(C)$ may serve as a functional of comparison.

Instead of postulating the uniform boundedness of the *lengths* of all curves C for which $J(C) \leq M$, it is sufficient to assume the uniform boundedness of *the values of some line integral of comparison* $L(C)$ for all curves C for which $J(C) \leq M$. Clearly this condition is weaker than the ordinary postulate.

10. GENERAL ANALYSIS AND THE EXISTENCE OF MINIMIZING CURVES

Before starting the discussion of generalized spaces we mention some consequences concerning the existence of minimizing curves in the ordinary Euclidean spaces resulting from the theorems mentioned in the preceding sections. In deriving from the lower semicontinuity of the integral the existence of curves minimizing it, we follow Tonelli's method, that is to say, we apply a theorem of the general analysis generalizing the classical theorem of R. Baire that a lower semicontinuous function on a compact domain of definition assumes its minimum. In fact, all the procedures of the calculus of variations since Hilbert's famous paper on the problem of Dirichlet which are known as "direct *methods*" are applications of *theorems* of the general analysis.

However it seemed desirable to formulate quite explicitly this theorem which implicitly underlies Tonelli's deductions. Only in this way was it possible to remove the length from its unjustified role as the only functional of comparison, and to introduce as a functional of comparison any lower semicontinuous functional of curves for which Hilbert's theorem holds. We call the theorem of the general analysis so obtained[20] "Tonelli's Principle." It reads as follows:

Let L be a limit class, i.e., a set of elements for which certain sequences of elements are said to converge toward a limit element. (In particular, each sequence all of whose elements are equal, is said to converge toward this element; moreover it is assumed that each subsequence of a sequence converging toward an element converges toward this element.)

Let β be a function defined in L, associating a (finite or infinite) number $\beta(e)$ with each element e of L, and let α be a function defined on a subset L_α of L containing, in particular, all elements e for which $\beta(e) < \infty$.

Let K be any subset of L_α closed in the set of all e for which $\beta(e) < \infty$.

If α and β satisfy the conditions

(1) For each number $\alpha_0 < \infty$ the function β is uniformly bounded above on the set of all elements e for which $\alpha(e) \leq \alpha_0$.

(2) For each $\beta_0 < \infty$ the set of all elements e for which $\beta(e) \leq \beta_0$, is compact,

then K contains at least one element e_0 that is limit of a minimizing sequence for α on K, i.e., a sequence $e_1, e_2, \cdots$ for which the numbers $\alpha(e_1), \alpha(e_2), \cdots$ converge toward the greatest lower bound of α on K.

If in addition to (1) and (2) the functions α and β satisfy the condition

(3) For each $\beta_0 < \infty$ the function α is lower semicontinuous on the set of all elements e for which $\beta(e) \leq \beta_0$,

then each element e which is limit of a minimizing sequence for α on K actually minimizes α on K.

Tonelli and his successors apply the special case of this theorem in which the elements of L are continuous curves of a Euclidean region, $\beta(e)$ is the length of the curve e, and $\alpha(e)$ the line integral in consideration formed along the curve e.

In the same way we apply the general theorem to any line integral $\alpha(e)$ satisfying the semicontinuity condition (3), and any functional of comparison $\beta(e)$ satisfying the analogue of Hilbert's theorem about the length expressed in condition (2), provided that α and β are connected by the condition (1). In this way our theorems about the semicontinuity of integrals enable us to prove in Tonelli's way the existence of curves minimizing the integral.

II. METRIC METHODS AND THE UNDERLYING SPACE

The classical theory is restricted to curves contained in a Euclidean space, or at any rate, to curves contained in a space whose points are defined by real coordinates. The integral is the variational length or F-length derived from the variational or F-distance, where var. dist. $pq = F(p, \theta_{pq}) \times$ Eucl. dist. pq.

Our theory can be developed[21] in any metric space in the sense of Fréchet. That is to say, we merely require a set of points such that with any two points a number is associated, called their geometric distance, satisfying the following conditions:

geom. dist. $pq =$ geom. dist. qp (Symmetry)
geom. dist. $pq > 0$ if $p \neq q$ (Positivity)
geom. dist. $pp = 0$ (Normality)
geom. dist. $qr +$ geom. dist. $rs \geq$ geom. dist. qs (Triangular inequality).

The points need not be defined by coordinates, nor need the distance be expressed in terms of coordinates.[22]

We have now to say what corresponds in our metric space to the integrand $F(p, \theta)$ depending upon a point and a direction in the Euclidean case. We assume to be given a function $F(p; q, r)$ of a point p and an ordered pair of points q, r which has the value 0 if $q = r$, and we set

var. dist. pq or F-dist. $pq = F(p; pq) \times$ geom. dist. pq.

From this point on we proceed as we did in the Euclidean case. We form the F-length of a polygon, and call F-length of a continuous curve contained in the metric space the limit which the F-lengths of the inscribed polygons approach when their norms approach 0, provided that such a limit exists.

The question is under what conditions concerning F and C the F-length of a curve C exists and is a lower semicontinuous functional. In other words, what assumptions concerning $F(p; q, r)$ defined in the metric space correspond to the continuity and quasi-regularity assumptions concerning $F(p, \theta)$ defined in the Euclidean space?

Clearly the continuity and semicontinuity properties of F with regard to p can be formulated if p is a point of any general metric space. But the metric methods enable us also to express quasi-regularity properties of an integrand $F(p; q, r)$ defined in a general metric space.

The easiest way to realize this possibility starts with a second interpretation of quasi-regularity of a function $F(p, \theta)$ in the Euclidean case. First we had considered convexity properties of the figurative. The second interpretation is given in terms of the *indicatrix* of F introduced by Carathéodory. On each ray θ issuing from the point p_0 we lay off a segment of length $1/F(p_0, \theta)$. The set of all endpoints so obtained is called the indicatrix of F at p_0. In the plane the indicatrix is a curve, in the n-space a $(n-1)$-dimensional surface intersecting each ray issuing from p_0 exactly once. (If $F(p_0, \theta)=0$ for some θ, then we say that the indicatrix intersects the ray θ at its point at infinity.) Now it can be shown that in the case of a nonnegative F, quasi-regularity of F at p_0 is equivalent to convexity of the indicatrix of F at p_0. On the other hand, according to a well-known theorem of Minkowski's, an $(n-1)$-

dimensional surface intersecting each ray of a bundle exactly once, is convex if and only if a certain metric associated with the surface satisfies the triangular inequality. The distance between two points q, r according to this Minkowski metric associated with the indicatrix $F(p_0, \theta)$, is the number $F(p_0, \theta_{qr}) \times$Eucl. dist. qr. (In particular, each point of the indicatrix of F at p_0 has the distance 1 from p_0, in this metric.) Hence quasi-regularity of F at the point p_0 is equivalent to the triangular inequality

$$\begin{aligned}&F(p_0, \theta_{qr}) \text{ Eucl. dist. } qr + F(p_0, \theta_{rs}) \text{ Eucl. dist. } rs\\ &\geq F(p_0, \theta_{qs}) \text{ Eucl. dist. } qs.\end{aligned}$$

F. Alt succeeded in proving this equivalence for any integrand $F(p, \theta)$, even for the case that $F(p_0, \theta)$ is positive for some directions θ and negative for others.[23] In this case the indicatrix has to have a property called projective convexity which for a positive function F is the same as the ordinary convexity.

Now such a triangular inequality can be formulated for a function $F(p; q, r)$ defined in any metric space. We postulate

$$\begin{aligned}&F(p_0; q, r) \text{ geom. dist. } qr + F(p_0; r, s) \text{ geom. dist. } rs\\ &\geq F(p_0; q, s) \text{ geom. dist. } qs\end{aligned}$$

for any four points p_0, q, r, s. Whereupon we can prove existence and semicontinuity of the variational length of a curve in the metric space exactly in the same way as we did in the Euclidean space.

It should be noted that the triangular inequality we had to postulate does not refer to the variational distance of points. By var. dist. qr or F-dist. qr we meant the number $F(q; q, r) \times$geom. dist. qr, whereas our assumed triangular inequality deals with the numbers $F(p_0; q, r) \times$geom. dist. qr, etc. Only if $q = p_0$ this number is the variational distance between q and r. For each point p_0 the number $F(p_0; q, r) \times$ geom. dist. qr gives rise to a metric satisfying the triangular inequality. We might call this metric "tangential" to the variational metric at the point p_0.

The variational length itself does not necessarily satisfy the triangular inequality, nor is it necessarily symmetric or positive. The only one of the four postulates of Fréchet it does satisfy is the normality condition

$$\text{var. dist. } pp = 0.$$

In addition, it has those properties which result from the semicontinuity of F, the triangular inequality for the tangential metrics and the properties of the underlying geometric distance. These assumptions are sufficient to prove the existence and semicontinuity of the integral and the existence of minimizing curves to the extent we developed the theory in the Euclidean case.[24]

There is another way of looking upon the situation. The theory of length had been developed not only for the curves contained in Euclidean spaces, but far beyond; e.g., the writer had studied the length of curves contained in general metric spaces in the sense of Fréchet.[12] All that actually was needed was to even omit some of the postulates of a general metric space, few as they are, and still develop a theory of length. A development of the theory of length under weaker and weaker conditions was bound to result in a theory general enough to comprise what we called variational length of curves. This program has been carried out in a recent publication.[25]

12. UPPER SEMICONTINUITY AND MAXIMUM PROBLEMS. CONTINUITY

We saw that the integral of a function F along a curve C, in the sense of the F-length of C, is lower semicontinuous on each set of curves of uniformly bounded lengths, if the integrand F is lower semicontinuous and positively quasi-regular. (Metric methods allow us even to admit 0-sets of exceptional points.) Positive quasi-regularity of F at the point $p_0=(x_0, y_0)$ means that the figurative of F at p_0 is convex downward, or, what is equivalent, that the indicatrix of F at p_0 is convex, or, in still other words, that the following triangular inequality holds

$$F(p_0, \theta_{qr})\times\text{Eucl. dist. } qr + F(p_0, \theta_{rs})\times\text{Eucl. dist. } rs \geq F(p_0, \theta_{qs})\times\text{Eucl. dist. } qs.$$

It goes without saying that the integral of F is upper semicontinuous on each set of curves of uniformly bounded lengths if the integrand F is upper semicontinuous and negatively quasi-regular. By negative quasi-regularity of F at the point p_0 we mean that the figurative of F at p_0 is convex upward, the indicatrix of F at p_0 concave, and the following counter-triangular inequality holds

$$F(p_0, \theta_{qr})\times\text{Eucl. dist. } qr + F(p_0, \theta_{rs})\times\text{Eucl. dist. } rs \leq F(p_0, \theta_{qs})\times\text{Eucl. dist. } qs.$$

It is further obvious that with each theorem about the existence of minimizing curves derived from the lower semicontinuity of the integral there corresponds a theorem about the existence of maximizing curves derived from the upper semicontinuity of the integral.

Continuity of the integral is equivalent to its being both lower and upper semicontinuous. Sufficient for the continuity of the integral on each set of curves of uniformly bounded lengths is thus that the integrand is both lower and upper semicontinuous, and both positively and negatively quasi-regular. This means that the integrand is continuous and its figurative at each point is flat, its indicatrix at each point is flat, and the following triangular equality holds for each p_0

$$F(p_0, \theta_{qr}) \times \text{Eucl. dist. } qr + F(p_0, \theta_{rs}) \times \text{Eucl. dist. } rs = F(p_0, \theta_{qs}) \times \text{Eucl. dist. } qs.$$

In the two-dimensional case of an integrand $F(x, y, u, v) = F(p, \theta)$ the figurative at the point (x_0, y_0), that is the surface $z = F(x_0, y_0, x', y')$, is flat (i.e., a plane through the origin of the (x', y', z)-space) if and only if for two constants u_0 and v_0 we have $F(x_0, y_0, x', y') = u_0x' + v_0y'$. Then the indicatrix has the following equation in polar coordinates: $\rho = 1/F(x_0, y_0, \cos\theta, \sin\theta) = 1/(u_0\cos\theta + v_0\sin\theta)$ which is the equation of a straight line.

Sufficient for the continuity of the integral on each set of curves of uniformly bounded lengths is thus that the integrand $F(x, y, \cos\theta, \sin\theta)$ is a continuous function of the form $u(x, y)\cos\theta + v(x, y)\sin\theta$, which is the case if $u(x, y)$ and $v(x, y)$ are continuous functions. The integral of this integrand along the curve C is usually written in the form $\int_C u(x, y)dx + v(x, y)dy$ or, briefly, $\int_C udx + vdy$.

13. LINE INTEGRALS INDEPENDENT OF THE PATH

We now turn to problem IV of section 2: Under what conditions is a line integral independent of the path, i.e., when does it assume the same value for any two coterminal curves of finite length? Such an integral clearly is a continuous functional. We saw in the last section that it is sufficient for the continuity of an integral that it be of the form $\int u(x, y)dx + v(x, y)dy$, where u and v are continuous functions. This condition is not necessary for the continuity of the integral. But although integrals of the form $\int udx + vdy$ are not the only continuous integrals, they seem to be the most interesting ones. Thus, as it is customary, we shall restrict the research to a study of the conditions under which an integral of the form $J(C)=\int_C u\,dx+v\,dy$ is independent of the path C. These integrals have the further important property that $J(-C)=-J(C)$, where $-C$ denotes the curve C traversed in the opposite direction.

If the first classical condition mentioned in section 3 holds, that is to say, if $\partial u/\partial y$ and $\partial v/\partial x$ exist, are continuous, and are equal for each (x, y) of a closed and bounded domain D in the (x, y)-plane, then the functions $\partial u/\partial y$ and $\partial v/\partial x$ are uniformly continuous, and for each $\epsilon>0$, one readily proves: There is a number $l_\epsilon>0$, such that if

$$(\underline{a}, \underline{b}),\ (\underline{a}, \bar{b}),\ (\bar{a}, \underline{b}),\ (\bar{a}, \bar{b})$$

are the vertices of any rectangle whose diagonal is $<l_\epsilon$, then the following inequality, to which we shall refer as the *fundamental ϵ-inequality*, holds

$$\left|\frac{u(\underline{a}, \bar{b})-u(\underline{a}, \underline{b})}{\bar{b}-\underline{b}}-\frac{v(\bar{a}, \underline{b})-v(\underline{a}, \underline{b})}{\bar{a}-\underline{a}}\right|<\epsilon$$

or, what is equivalent,

$$|[v(\underline{a},\underline{b})(\bar{b}-\underline{b})+u(\underline{a},\bar{b})(\bar{a}-\underline{a})]-[u(\underline{a},\underline{b})(\bar{a}-\underline{a})+v(\bar{a},\underline{b})(\bar{b}-\underline{b})]|< \epsilon(\bar{a}-\underline{a})(\bar{b}-\underline{b}).$$

Now let R be the rectangle $\underline{x}\leq x\leq\bar{x}$, $\underline{y}\leq y\leq\bar{y}$. By a rectangular net in R, we mean a matrix of points (x_i, y_j) $(i=0, 1, \cdots, m;\ j=0, 1, \cdots, n)$ for which

$$x_0=\underline{x},\ x_m=\bar{x},\ x_i<x_{i+1};\ y_0=\underline{y},\ y_n=\bar{y},\ y_j<y_{j+1}.$$

Each rectangle

$$(x_i, y_j),\ (x_{i+1}, y_j),\ (x_i, y_{j+1}),\ (x_{i+1}, y_{j+1})$$

is called a mesh of the aforementioned net, N. The largest diagonal of a mesh, i.e., the largest of the numbers $\sqrt{(x_{i+1}-x_i)^2+(y_{j+1}-y_j)^2}$, will be called the norm of the net N.

Now let us denote by R_1 and R_2 the two curves contained in the boundary of R, starting at the point $(\underline{x}, \underline{y})$, and terminating at the point $(\bar{x}, \bar{y})$. Let R_1 consist of the left and upper, R_2 of the lower and the right side of R. We shall set

$$J(R_1) = \int_{R_1} u\,dx + v\,dy \text{ and } J(R_2) = \int_{R_2} u\,dx + v\,dy.$$

Let us furthermore denote by $T_1(N)$ and $T_2(N)$ the ordered sets of points of the net N on R_1 and R_2.

$$T_1(N)\colon (\underline{x}, y_0), (\underline{x}, y_1), \cdots, (\underline{x}, y_{n-1}), (\underline{x}, \bar{y}), (x_1, \bar{y}), \cdots, (x_m, \bar{y})$$
$$T_2(N)\colon (x_0, \underline{y}), (x_1, \underline{y}), \cdots, (x_{m-1}, \underline{y}), (\bar{x}, \underline{y}), (\bar{x}, y_1), \cdots, (\bar{x}, y_n)$$

Obviously, the norms of $T_1(N)$ and $T_2(N)$ are smaller than the norm of the net N.

Now let us assume that for some $\epsilon > 0$, the fundamental ϵ-inequality holds for each sufficiently small rectangle contained in the rectangle R. Then it is easy to see that the Weierstrass sums for the integrals $J(R_1)$ and $J(R_2)$, associated with the sets $T_1(N)$ and $T_2(N)$, differ by less than $2\epsilon(\bar{x} - \underline{x})(\bar{y} - \underline{y})$. Consequently, if for each $\epsilon > 0$ the fundamental ϵ-inequality holds for each rectangle contained in R, then the Weierstrass sums of $J(R_1)$ and $J(R_2)$ associated with sets of arbitrarily small numbers differ by as little as we please, and, hence, $J(R_1) = J(R_2)$, from which it easily follows that, within the rectangle R, the integral $J = \int u\,dx + v\,dy$ is independent of the path.

An analysis of the outlined proof shows[26] that the crucial equality $J(R_1) = J(R_2)$ can be derived from a weaker assumption than the condition that the fundamental ϵ-inequality holds for each sufficiently small rectangle contained in R. It is entirely sufficient to know that for each ϵ there is one rectangular net in R whose norm is less than ϵ, and each of whose meshes satisfies the fundamental ϵ-inequality. This condition is much weaker than the first classical assumption, from which it can be derived. It holds even in many cases in which neither u nor v have any partial derivatives whatever. For instance, it holds in the case of the integral $\int f(x+y)\,(dx+dy)$, where f is a continuous, nowhere differentiable function of one real variable—the case discussed in section 5, which cannot be covered by the classical theory since the integrand does not admit partial derivatives. For this integral, the fundamental ϵ-inequality is satisfied with regard to each square whose sides are parallel to the axes. If

$$(a, b),\ \ (a, b+s),\ \ (a+s, b),\ \ (a+s, b+s)$$

are the corners of such a square, then the left side of the fundamental ϵ-inequality for this square is

$$|\ [f(a+b)s+f(a+b+s)s]-[f(a+b)s+f(a+s+b)s]\ |.$$

Since this expression is obviously equal to zero, we see that for each square whose sides are parallel to the axis, the fundamental ϵ-inequality holds even for each ϵ. Consequently, the fundamental ϵ-inequality holds for each mesh of a net all of whose meshes are squares, and since there are such quadratic nets of arbitrarily small norms, it is clear that our sufficient condition holds, which guarantees that the integral is independent of the path.

The sufficient condition for the independence of the integral that we mentioned above is, however, not necessary. If we consider the following integral

$$\int |\ (x-y)\ |\ (dx-dy),$$

then it is clear that the integral is independent of the path and yet no rectangle whose left lower corner is the origin satisfies the fundamental ϵ-inequality for any number $\epsilon < 2$. Consequently, if we denote by R the unity square in the first quadrant, then there is no rectangular net in R all of whose meshes satisfy the fundamental ϵ-inequality for sufficiently small values of ϵ.

However by considering what we call dotted nets,[26] instead of nets, we can formulate a condition that is both sufficient and necessary in order that the integral $\int udx + vdy$ be independent of the path. By a dotted net we mean a rectangular net on the boundaries of whose meshes finite ordered sets of points are inserted. Let

$$(\underline{a}, \underline{b}),\ (\bar{a}, \underline{b}),\ (\underline{a}, \bar{b}),\ (\bar{a}, \bar{b})$$

be the vertices of a rectangle M and let

$$a_0 = \underline{a} < a_1 < \cdots < a_{m-1} < a_m = \bar{a};$$
$$b_0 = \underline{b} < b_1 < \cdots < b_{n-1} < b_n = \bar{b}.$$

By the dotted rectangle M, we mean the four vertices of M together with the ordered set of points $(\underline{a}, b_1)$, $(\underline{a}, b_2)$, $\cdots$, $(\underline{a}, b_{n-1})$ on the left side, and three corresponding ordered sets on the other three sides of M. We shall say that this dotted rectangle M satisfies the strong ϵ-inequality if

$$\Big|\Big[\sum_{j=0}^{n-1} v(\underline{a}, b_j)(b_{j+1} - b_j) + \sum_{i=0}^{m-1} u(a_i, \bar{b})(a_{i+1} - a_i)\Big] -$$

$$\Big[\sum_{i=0}^{m-1} u(a_i, \underline{b})(a_{i+1} - a_i) + \sum_{j=0}^{n-1} v(\bar{a}, b_j)(b_{j+1} - b_j)\Big]\Big|$$

$$< \epsilon(\bar{a} - \underline{a})(\bar{b} - \underline{b}).$$

In order that the integral $\int udx + vdy$ be independent of the path, it is necessary and sufficient that for each ϵ, there exist a dotted net whose norm is less than ϵ and each of whose dotted meshes satisfies the strong ϵ-inequality. In a way, our necessary and sufficient condition seems to be not so different from the statement that the integral is independent of the path, but this is only apparent. In order to understand the great difference between our condition and the independence statement, one has to realize that our condition refers for each ϵ to only one net whose norm is less than ϵ and which satisfies the strong ϵ-inequality. How weak this condition is can be seen from the fact that it is much weaker than the sufficient condition concerning undotted nets that we had studied before, which in turn was much weaker than the first classical assumption, since it took care of cases in which the integrand does not have any partial derivatives. By considering one sequence of dotted nets, the k-th of which has a norm less than $1/k$ and satisfies the strong $1/k$-inequality, we reach a condition still sufficient, but so weak that it is also necessary. It should be noticed that this condition can be directly inferred[26] also from the second classical condition mentioned in section 3, the Goursat-Heffter condition dealing with functions u and v that have differentials, although it is hardly necessary to remark that on the other hand our condition is much weaker than also the second classical assumption. In a recent paper Fubini[27] formulated a necessary and sufficient condition of the type of the one presented above but dealing with a simpler concept of dotted nets.

What has this theory to do with metric geometry? It is clear from what we said that our conditions can be formulated without referring to metric concepts. But it is also clear that the introduction of a metric terminology is helpful and suggestive, and that the leading ideas of our theory are suggested by the methods of metric geometry.

The fundamental ϵ-inequality for a rectangle simply means that the variational polygonal length of the polygon R_1, and the variational polygonal length of the polygon R_2 differ by a quantity that is small compared with the area of R. Here R_1 is the polygon consisting of the left lower, the left upper, and the right upper vertices of R; and R_2 is the polygon consisting of the left lower, the right lower and the right upper vertices of R. The strong ϵ-inequality for a dotted rectangle states that the variational polygonal lengths of two polygons containing R_1 and R_2, respectively, have a difference which is small compared with the area. It is interesting and instructive to realize the geometric meaning of each step of the proofs which, in fact, were found by such geometric considerations.

14. RÉSUMÉ

Owing to the limitations of a lecture, our exposition of the metric methods in analysis must necessarily be incomplete in many ways.

More results can be obtained along these geometric lines than could be mentioned here. It is clear that what we have said about integrals of real functions independent of the path has a bearing on the theory of functions of one or more complex variables. Furthermore, metric methods yield new proofs for the classical theorems of vector analysis, in most cases under conditions weaker than the classical ones. Finally, our methods can be applied to multiple integrals.

All we could cover in this lecture was the enumeration of some of the results, restating classical theorems under weaker conditions. Taken separately, some of these generalizations may seem less significant. What one should have in mind is that our metric methods lead to progress concerning line integrals in many directions at the same time. (Naturally we did not even mention progress due to these methods in other fields of analysis, as, for instance, in differential geometry, of which a résumé is to be found in L. M. Blumenthal's book, *Distance Geometries*.[12])

Another aspect which may be as important as the generalizations of the classical results is the modification of their proofs due to the introduction of metric methods, and this point naturally had to be omitted in a single lecture altogether; but an analyst, studying the metric proofs of the generalized classical results, will notice considerable deviations from the classical proofs.

One of the characteristics of metric proofs is that one can visualize each step of the argument. What these methods try to introduce into analysis is a complete *geometrization*. With regard to the existence theorems of the calculus of variations, this aim has been accomplished. It is a well known fact that some analysts are not too fond of geometric methods. (Twenty years ago, an eminent analyst was reported to have defined geometry as that branch of mathematics in which false statements are called evident.) If one does not like geometric intuition accompanying analytic arguments, then one need not evoke the visualization of the metric arguments in analysis, but if one does like it one may.

Another characteristic of metric proofs is the *logicalization* of the theory which is due to them. They try to derive statements as comprehensive as possible from a minimum of assumptions, in the simplest way. Only such hypotheses are assumed as are actually used, and the proofs make it clear where and to what extent the assumptions are used. Metric solutions of problems of analysis are of the type of those postulational theories in which the arguments are interlocked like the wheels of a watch of greatest efficiency and simplicity. In particular, they do not follow the century-old use of starting with differentiability assumptions—of starting with them in order to be able to apply the calculus. (The legend does not say whether a geometrist retorted to the analyst's definition of geometry, that analysis is that branch of mathematics which derives interesting conclusions from superfluous assumptions!)

It is one of the wonderful achievements of calculus that for many individual curves which are given by simple functions, we can find the length by simple processes of differentiation and integration, the differentiation applicable only to curves with tangents. E.g., we can say right away that the arc of the parabola given by the equation $y=x^2/2$ $(0 \leq x \leq a)$ has the length

$$\int_0^a \sqrt{1+x^2}dx = \frac{a}{2}\sqrt{1+a^2} + \frac{1}{2}\log(a+\sqrt{1+a^2}).$$

Statements of this type which are of the highest importance will always remain the domain of calculus. But for certain purposes, we are also interested in statements concerning length which are of another type, namely general statements about properties of the length of all curves, e.g., that the length is lower semicontinuous, that is to say, that all curves sufficiently close to a curve C of finite length can not be much shorter than C. Why should we, in proving a theorem of this type which holds for *all* curves, restrict ourselves to the consideration of curves whose lengths can be computed by certain simple processes, important as this computation may be for those curves to which it is applicable? Why should we artificially base the study on assumptions necessary for the application of calculus but without any bearing on the problem? The enormous successes of calculus in its proper domain should not induce us to apply its ideas beyond this proper domain. While computations of quantitative properties of many simple individual objects belong to the domain, many comprehensive statements about all objects of more general type seem to lie beyond the scope of classical analysis. Classical analysis frequently can obtain general statements of great heuristic value, but in some cases not the definite answers.

The metric attack on such problems is a very direct one, admitting only the restrictions necessitated by the problem, without introducing limitations which are due to the method.

NOTES

[1] Short proofs of these theorems are to be found, e.g., in Tonelli's *Fondamenti di Calcolo delle Variazioni*, Bologna, 1921, and in Saks's *The Theory of the Integral*, New York, 1937.

[2] The most thorough investigations of these questions seem to be contained in the papers by Pringsheim in the *Sitzungsber. Akad. Wiss., München*, (1895–1903).

[3] *Math. Ann.*, LXIII (1907), p. 254.

[4] *Duke Math. Jour.*, Vol. 4 (1938), p. 132.

[5] See *Ergebnisse e. math. Kolloquiums*, Vienna, 8, 1937, p. 17 (footnote).

[6] *Ann. Ec. Norm.*, t. 24 (1907).

[7] *Acta Litt. Sci.*, Szeged, IV (1929), p. 38.

[8] *Variationsrechnung*, Leipzig, 1935, p. 310.

[9] See the author's paper in the *Proc. Nat. Acad. Sci.*, Vol. 25 (1939), p. 621.

[10] *Essai sur l'unité des méthodes directes*, Bruxelles, Hayez, 1933, p. 32.

[11] *Fundamenta Math.*, XXV (1935), p. 441; *Comptes Rendus*, Paris, t. 201 (1935), p. 705; t. 202 (1936), p. 1007, p. 1648; *Proc. Nat. Acad. Sci.*, Vol. 23 (1937), p. 244; Vol. 25 (1939), p. 621; and the comprehensive paper, "Metrische Geometrie und Variationsrechnung," *Ergebnisse e. math. Kolloquiums*, 8, 1937, p. 1.

[12] In particular, the paper in the *Math. Ann.*, CIII (1930), pp. 466–501. A survey of the main results of metric geometry and its applications up to the year 1938 is contained in L. M. Blumenthal's book, *Distance Geometries: A Study of the Development of Abstract Metrics*, The University of Missouri Studies, Vol. 13 (1938). See also the author's paper, "La géométrie des distances et ses relations avec les autres branches des mathématiques," *L'Enseignement Math.*, t. 35 (1936), p. 348. Interesting mathematical and historical remarks are to be found in a paper by Pauc in the *Ann. Sc. Norm. Sup.*, Pisa, VIII (1939), concerning what Pauc calls the Weierstrass-Bouligand-Menger integral and its relations to the Riemann and Lebesgue integrals.

[13] *Ergebnisse e. math. Kolloquiums*, 8, 1937, p. 34.

[14] See the definition of $\tau(p, \rho)$, *loc. cit.*, p. 11 (§11), and the assumptions III and IV, p. 14 and p. 28.

[15] *Loc. cit.*, p. 19 (§ 28) and the theorems on p. 20.

[16] See the condition V* on p. 29 of the paper, *loc. cit.*

[17] *Loc. cit.*, p. 20 (§ 31).

[18] *Loc. cit.*, Chapter VI, pp. 29–32.
[19] *Loc. cit.*, pp. 21–23.
[20] *Loc. cit.*, Chapter II, p. 13, and *Proc. Nat. Acad. Sci.*, Vol. 25 (1939), p. 621.
[21] *Loc. cit.*, and *Proc. Nat. Acad. Sci.*, Vol. 23 (1937), p. 244.
[22] The results may be applied to Hilbert space. See Wald *loc. cit.* note 13, p. 36.
[23] *Loc. cit.*, note 13, p. 32.
[24] *Proc. Nat. Acad. Sci.*, Vol. 23 (1937), p. 244.
[25] *Loc. cit.*, Vol. 25 (1939), p. 621.
[26] *Proc. Nat. Acad. Sci.*, Vol. 25 (1939), p. 621.
[27] *Proc. Nat. Acad. Sci.*, Vol. 26 (1940), p. 199.

II

ON ALGEBRA OF GEOMETRY AND RECENT PROGRESS IN NON-EUCLIDEAN GEOMETRY

I. THE CLASSICAL FOUNDATIONS OF GEOMETRY

IN 1927 a lecturer on the foundations of geometry used to follow one of two courses: Hilbert's way[1], starting with five groups of assumptions about a large number of undefined concepts and relations (point, line, plane, incidence relations, an order relation, congruence relations for segments and for angles, etc.)—or Veblen's method, basing projective geometry[2] on postulates about two undefined concepts (point and line) and one relation (a point lies on a line), and starting affine geometry[3] with only the concept of point and an order relation (a point lies between two points).

Other theories had been developed. Let us mention only Pieri's profound papers published at the turn of the century[4], and Huntington's remarkable foundation of Euclidean geometry[5] in terms of the concept of "sphere" and the relation "contained in." While the latter method does not seem to be applicable to projective and affine geometry, and thus left Euclidean geometry rather isolated, Pieri's results were developed in Peano's logistic symbolism which, especially at the time of their publication, must have made them difficult reading for the majority of geometers. This may be one of the reasons why their mathematical content was not given the full attention it deserved, and the two ways mentioned at the outset became best known.

Hilbert's development, however, is rather cumbersome. Veblen's method bases the self-dual system of projective geometry on the two concepts of point and line which, except for the case of the plane, are non-dual.

But there were more serious objections. In developing four-dimensional geometry along Hilbert's line, one has to introduce a fourth undefined concept, that of "three-dimensional flat," and for each higher dimension an additional concept. Moreover, for each higher dimension, one has to introduce new postulates concerning these higher dimensional concepts. Veblen, starting with two basic classes of undefined entities, called points and lines, introduces planes (and, in higher dimensional spaces, higher dimensional flats) as sets of points. This is not only an asymmetry contrasting with the symmetric role played by points, lines, and planes in projective geometry, but, strictly speaking, makes Veblen's foundation of projective geometry dependent upon large parts of the theory of sets. Each theorem about all planes or about the existence of certain planes in this theory is a universal or existential proposition concerning the system of all subsets of the basic classes.

Moreover, a lecturer on geometry in 1927 missed in both standard foundations of geometry the extreme rigor and completeness to which he was accustomed in the foundations of Boolean algebra, topology, and algebra. He did not find among the postulates a statement like "the line determined by the points P and Q, and the line determined by the points Q and P, are identical," although this simple assumption was used throughout the development of geometry. In algebra, one explicitly mentioned among the postulates the statement that the sum of the numbers a and b, and the sum of the numbers b and a, are equal, for the reason that in the development of algebra the formula $a+b=b+a$ was used.

In a similar way, the analogue of the associative law of addition, which is among the assumptions of algebra, was not mentioned among the assumptions of geometry although it was constantly used. In Pieri's theory complete rigor was observed, but, as we mentioned, at the expense of simplicity of reading. Not less than four or five of his postulates translated from the logistic symbolism into ordinary language have merely the content that there exist at least three non-collinear points.

Finally, the problem of a comprehensive or simultaneous postulational treatment of projective, affine, and Euclidean geometry was not solved in an entirely satisfactory way. Hilbert's grouping of his assumptions for Euclidean geometry into a group of postulates dealing merely with the projective incidence relation, a group of axioms of order, a group of congruence axioms, etc., suggests the possibility of developing first the theorems common to Euclidean, affine, and projective geometry, then the theorems common to Euclidean and affine geometry, and finally the typically Euclidean theorems. But Hilbert does not really carry out this program. Veblen bases projective and affine geometry not only on different postulates—this is, of course, unavoidable since they are different theories—but, as we mentioned, on different undefined concepts. Now one might say that undefined concepts about which we make assumptions that are essentially different (i.e., lead to non-isomorphic theories) are always different, since the content of the undefined concepts merely consists in the assumptions about the concepts. But it remained that the difference between the postulates for projective and for affine geometry was much greater than that between the two theories as a whole. Since projective and affine geometry have so much in common, why not base them on two sets of assumptions that have much in common?

The relations between projective, affine, and other geometries were usually described along the lines of Klein's Erlanger Program in terms of groups of transformations. In this sense the other geometries were subordinated to projective geometry. The postulational developments since Pasch[6] established the relation between projective and affine geometry in either of the following ways: The projective space is obtained from the affine space by adding to it an ideal plane and the points and lines of this plane ("at infinity"). The affine space is obtained from the projective space by omitting a plane and the points and lines of this plane ("at infinity").

But the parallelism between projective and affine geometry mentioned above, suggested the possibility of coordinating projective and affine geometry instead of subordinating the one under the other. And it was to be expected that such a parallel development of the two geometries from two sets of assumptions which have much in common, might lead to the coordination of further postulational theories, e.g., of non-Euclidean geometry. One could even hope that the minor differences between the sets of basic assumptions might constitute another principle of classifying geometries.

2. THE IDEA OF AN ALGEBRA OF GEOMETRY

The author of this paper when lecturing on the foundations of geometry in 1927 noticed that all the difficulties of the existent theories can be solved and all the suggestions to which their criticism leads can be carried out by one simple idea: that of starting with assumptions about only one basic class of undefined entities which we shall call "flats," this system of flats to comprise points, lines, planes, and in higher dimensional spaces also higher dimensional flats. By doing so, we avoid the necessity of adding new concepts and assumptions for higher dimensions. Moreover, we escape any asymmetry or non-duality of the basic concepts. Finally, we do not have to consider general subsets of our basic class in order to define higher dimensional flats. The universal and existential propositions of our theory are limited to pairs, triples, and quadruples of elements, and in a few cases to finite sets of elements of our class. Thus the set theoretical basis of our theory is very restricted.

It goes without saying that in such a theory the assumptions will have to be statements which hold for flats of each dimension. Since the more specific geometric statements hold only for flats of a certain dimension, we can thus expect that our assumptions will be of a very simple and fundamental nature—just of the type of those hypotheses which in the standard foundations were not mentioned at all. Since all our assumptions will have to be of a very simple and fundamental nature, we can expect furthermore that, if anything, then only the utmost completeness with regard to our assumptions can lead to the derivation of a geometric theory from them. Thus, the complete rigor observed in Boolean algebra, topology, or algebra will be a necessary feature of a theory of these flats. The question was how far one can get in geometry with assumptions which are so simple that they hold for flats of all dimension. It seemed worthwhile to try.

The algebra of numbers has been developed from postulates about adding and multiplying numbers; the algebra of classes from postulates about joining and intersecting classes. This suggested a foundation of geometry on postulates about joining and intersecting flats, and the name "algebra of geometry" for the theory so developed.

Later Ore pointed out[8] that for algebraic purposes Dedekind in his notes to Dirichlet's "Vorlesungen ueber Zahlentheorie" had developed a calculus of modules which, when applied to geometry, leads to a foundation of geometry on two operations. A module is a system of numbers which with any two numbers contains their difference. If M_1 and M_2 are two modules, then Dedekind denotes by M_1+M_2 the set of all numbers which are the sum of one number of M_1 and one number of M_2. He calls M_1-M_2 the set of all numbers which belong both to M_1 and M_2. In this way he clearly has $M-M=M$ and $M_1-M_2=M_2-M_1$. If we apply this idea to analytic geometry in which each point is given by n coordinates, and each flat determined by a system of homogeneous linear equations, then we see that each flat represents a module whose "numbers" are the points of the flat. For if two points $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_n$ lie on the flat, i.e., satisfy a system of linear equations, then the "difference point" $x_1-y_1, x_2-y_2, \cdots, x_n-y_n$ lies also on the flat. The sum of two flat modules is the flat of least dimension containing the two summand flats. Dedekind's "difference" of two flats is the flat of highest dimension contained in both flats; this is what we call the intersection of the two flats. While the algebra of modules can in this way be applied to geometry, Dedekind himself, as his notation shows, does not seem to have had in mind any geometric applications. He mainly dealt with the likewise interesting analogy between the sum of two modules and the greatest common divisor, and between the difference of two modules and the least common multiple. Ore applied these ideas to group theory.[8]

In this lecture, as in the series of my former publications on this subject, I shall present the foundation of geometry in terms of the two operations of joining and intersecting, denoted by + and · . But I should like to mention also another way followed independently by Garrett Birkhoff[9]. It is based on Hausdorff's concept of a partially ordered set, i.e., a set in which a relation, denoted by $\subset$, is given for some pairs of elements satisfying the following conditions

(1) For no element x we have $x \subset x$.
(2) If $x \subset y$, then y not $\subset x$.
(3) If $x \subset y$, and $y \subset z$, then $x \subset z$.

In other words, the $\subset$-relation in a partially ordered set is assumed to be irreflexive, asymmetric, and transitive like the $<$-relation defined in the set of all real numbers. But whereas the set of all real numbers is ordered, that is to say, for each pair of different numbers x and y, we have either $x<y$ or $y<x$, in a partially ordered set we only postulate that for some pairs of different elements x and y, we have $x \subset y$ or $y \subset x$, and admit the existence of different elements for which neither $x \subset y$ nor $y \subset x$. For instance, the complex numbers form a partially ordered set if for the two numbers $x = x_1 + ix_2$ and $y = y_1 + iy_2$ we set $x \subset y$, if and only if $|x| < |y|$ where $|x| = +\sqrt{x_1^2 + x_2^2}$ and $|y| = +\sqrt{y_1^2 + y_2^2}$.

In order to build up projective geometry, Birkhoff makes two postulates guaranteeing that for any two elements of the partially ordered set, A and B, there exists a join (i.e., an element X such that $A \subseteq X$ and $B \subseteq X$, and that $A \subseteq X'$ and $B \subseteq X'$ imply $X \subseteq X'$), and a meet (i.e., an element Y such that $Y \subseteq A$ and $Y \subseteq B$, and that $Y' \subseteq A$ and $Y' \subseteq B$ imply $Y' \subseteq Y$). Birkhoff in his extended theory calls a system satisfying these conditions a lattice.

Returning to our algebra of geometry, we see that it has n common with algebra of numbers and algebra of classes wo binary operations, denoted by + and · which are associative and commutative, and admit indifferent elements. Indifferent with regard to addition and multiplication are 0 and 1 among the numbers, the empty and the total class among the classes, the vacuous flat V and the universal flat U among the flats. For any flat A we have $V+A=U\cdot A=A$. Finally in all three algebras, the indifferent element of addition reproduces itself under multiplication; we have $n\cdot 0=0$ for each number, $A\cdot V=V$ for each class A and each flat A.

Another feature common to the algebras of numbers and of classes, the distributive law $A\cdot(B+C)=(A\cdot B)+(A\cdot C)$, is not postulated in the algebra of geometry. If B and C are distinct points, A a straight line intersecting their join $B+C$ without passing through either of the points B and C, then $A\cdot(B+C)$ is a point whereas $(A\cdot B)+(A\cdot C)=V$. The non-validity of the distributive law is one of the distinctive characteristics of algebra of geometry.

However, algebra of geometry has two features in common with the algebra of classes which do not hold in algebra of numbers. The first is the dual of the formula $V\cdot A=V$, i.e., $U+A=U$. For numbers, the equality $1+n=1$ is false unless $n=0$. The second is that both flats and classes are idempotent under both joining and intersecting, that is to say that $A+A=A\cdot A=A$. In the field of real and complex numbers, the only element which is idempotent with regard to an operation is the indifferent element of the operation.

3. THE ALGEBRA OF PROJECTIVE GEOMETRY

For projective geometry, the shortest foundation[10] is obtained by postulating associativeness of the two operations, the existence of the indifferent elements, either one reproducing itself under the other operation, and the following law which we shall call the weak distributive law:

$$A+[(A+B)\cdot C]=A+[(A+C)\cdot B],$$
$$A\cdot[(A\cdot B)+C]=A\cdot[(A\cdot C)+B].$$

By setting $A=V$ in the first, and $A=U$ in the second of these formulae, we see that both operations are commutative. By setting $C=V$ and $C=U$, respectively, we get $A=A+(A\cdot B)$ and $A=A\cdot(A+B)$. If in these last formulae we set $B=U$ and $B=V$, respectively, we derive $A=A+A$ and $A=A\cdot A$. Each element A is idempotent under the two operations. Thus there are neither multiples nor powers in the formulae of algebra of geometry. We express this fact by saying that the operations $+$ and $\cdot$ are *totally linear*. Furthermore, we see that if for two elements A and B we have $A+B=A$, then we also have $A\cdot B=B$, and conversely.

If for two elements A and B both formulae $A+B=A$ and $A\cdot B=B$ hold, then we call B a part of A and write $B\subseteq A$. If $B\subseteq A$ and $B\neq A$, then we call B a proper part of A, and write $B\subset A$. By introducing this relation which can be likewise introduced in the algebra of classes, we attain another analogy with the algebra of real numbers. For it is similar to the $<$-relation defined for real numbers. But whereas the $<$-relation makes the field of real numbers an ordered set, the $\subset$-relation makes our system of flats a partially ordered set. The $<$-relation is connected with the operations in the field of numbers by the laws: $m_1\leq m_2$ and $n_1\leq n_2$ imply $m_1+n_1\leq m_2+n_2$; and, if m_1 and n_1 are >0, then $m_1\cdot n_1\leq m_2\cdot n_2$. In the same way $A_1\subseteq A_2$ and $B_1\subseteq B_2$ imply $A_1+B_1\subseteq A_2+B_2$ and, without exception, $A_1\cdot B_1\subseteq A_2\cdot B_2$. In particular we have $V\subseteq A\subseteq A\subseteq U$ for each A.

An element which is different from V, and does not contain any parts besides itself and V, shall be called a point. An element which is different from U, and is not part of any element besides itself and U, shall be called a hyperplane. Our definition makes precise the famous first words of Euclid's elements: "Point is that which has no parts." The closest analogue in arithmetic is the concept of a prime number, i.e., an integer which is different from 1 and does not have any integer factor besides itself and 1.

As O. Schreiber showed in his Vienna thesis in 1933 one can deduce from our assumptions an important theorem about the insertion of an element between two other elements. It had been formulated as one of the basic postulates for a projective theory of dimension by G. Bergmann[11] in 1928. Since the word interpolation has another standard meaning, we shall refer to it as the Intercalation Law: If P is a point, and A and B are any two elements for which $A \subseteq B \subseteq A+P$, then $B=A$ or $B=A+P$. In the same way as the intercalation law we can derive its dual which formally can be obtained by multiplying the definition of a hyperplane H by B: If H is a hyperplane, and A and B are any two elements such that $A \supseteq B \supseteq A \cdot H$, then $B=A$ or $B=A \cdot H$.

A finite set of points $P_1, P_2, \cdots, P_m$ is called independent if no point of the set is a part of the sum of the other points of the set. (It should be noted that here for the first time the concept of a general finite set is introduced. So far we concerned ourselves with only pairs, triples, or quadruples of elements.) From the intercalation law we derive by means of complete induction the following fundamental theorem on independent points: If $P_1, P_2, \cdots, P_m$ are m independent points, and $Q_1, Q_2, \cdots, Q_n$ are n points (not necessarily independent) such that

$$P_1+P_2+\cdots+P_m=Q_1+Q_2+\cdots+Q_n,$$

then among the points $Q_1, Q_2, \cdots, Q_n$ exist m points Q_{i_1}, $Q_{i_2}, \cdots, Q_{i_m}$ such that

$$P_1+P_2+\cdots+P_m=Q_{i_1}+Q_{i_2}+\cdots+Q_{i_m}$$

(hence necessarily $n \geqq m$). If A is the sum of a finite set of points, then A is also the sum of an independent set of points, and the number of independent points with the sum A is the smallest number of points with the sum A. If two independent sets of points have the same sum, then they contain equally many points. In the same way one can deduce a dual fundamental theorem on independent hyperplanes.

These two theorems allow us to introduce two concepts of dimension. We shall define the additive dimension of an element A as one less than the number of independent points having the sum A, if A is the sum of a finite set of points; otherwise we call ∞ the additive dimension of A, if A is $\neq V$, and we call -1 the additive dimension of V. By virtue of the fundamental theorem on independent points, the number so defined depends upon A alone and does not depend upon the special way of representing A as the sum of an independent set of points. We define the multiplicative dimension of an element A as one less than the number of independent hyperplanes having the product A, if A is the product of a finite number of hyperplanes; otherwise we call ∞ the multiplicative dimension of A if A is $\neq U$, and we call -1 the multiplicative dimension of U.

Both kinds of dimension, the additive and the multiplicative, satisfy an important inequality: for any two elements A and B,

$$d(A)+d(B)\geqq d(A+B)+\dim(A\cdot B).$$

If we now make further assumptions to the effect that each element has a finite additive and a finite multiplicative dimension, then, as F. Alt and I proved[7], for any two elements, the sum of the additive and the multiplicative dimension is the same. Combined with the inequalities mentioned above, this leads to the fundamental dimension formula of projective geometry:

$$\dim A+\dim B=\dim(A+B)+\dim(A\cdot B).$$

While this formula holds both for additive and multiplicative dimension it is customary to restrict the term dimension to what we called additive dimension in the preceding paragraph.

4. SETS OF ELEMENTS. FUNCTIONS. COORDINATES

So far we have not referred to sets of elements beyond finite sets. For certain purposes it is of interest to study for a given element A the set of all points which are $\subseteq A$, and more general, for two given elements A and B and three given numbers a, b, n the set of all n-dimensional elements whose intersections with A and B have the dimensions a and b, respectively. Also the dual concepts are of interest.

The main importance of such sets lies in the fact that they are the domain of definition of interesting functions. E.g., the perspective transformation of the straight line L on the coplanar straight line M from the center of perspectivity O is nothing but the function $(X+O)\cdot M$ defined on $[L]_0$, the set of all points $\subseteq L$, i.e., for each X such that $V\subset X\subset L$. The projective transformation of L on N obtained by first applying the perspectivity on M from the center O, and then the perspectivity on N from the center P, is the function $[(X+O)\cdot M+P]\cdot N$ with the domain of definition $[L]_0$.

The algebra of geometry allows us to establish an algebra of functions of this type. The situation is somewhat similar to that in analytic geometry where a point is defined as an n-tuple of coordinates, a flat as the set of all points satisfying certain linear equations, and where an algebra of flats can be established dealing merely with the coefficients of the equations.

Projective geometry has not as yet been fully developed on the basis and in the notation of algebra of geometry. Therefore, I should like to mention a few examples illustrating some of the advantages and disadvantages of such a development.

The laws of Desargues and Pappus can, of course, be translated into formulae. The latter reads as follows: If $\dim P_i = 0$,

$$P_i \neq (P_1+P_3+P_5)\cdot(P_2+P_4+P_6) \qquad (i=1,\ 2,\ 3,\ 4,\ 5,\ 6)$$

$$P_1 \neq P_3 \neq P_5 \neq P_1, \quad P_2 \neq P_4 \neq P_6 \neq P_2,$$

$$\dim(P_1+P_3+P_5)=1, \ \dim(P_2+P_4+P_6)=1,$$

then $\dim[(P_1+P_2)\cdot(P_4+P_5)+(P_2+P_3)\cdot(P_5+P_6)+(P_3+P_4)\cdot(P_6+P_1)]=1.$

There are formulae holding for flats of any dimension, e.g., the weak distributive law

$$A+[(A+B)\cdot C]=A+[(A+C)\cdot B].$$

So far, no formulae have been discovered generalizing the laws of Pappus and Desargues and holding for flats of any dimension. We thus have to express the laws by rather cumbersome implications deriving formulae from assumptions about the dimension of the flats involved in the formulae. Complete rigor is attained in this way but as yet without advantages comparable to those which we derived from the algebraic viewpoint in the preceding sections.

As another more interesting example, we study the introduction of coordinates on a straight line contained in a projective plane. Let L be a straight line, A, B, C, three points of L such that $A \neq B \neq C$ (the possibility of $A=C$ is not excluded). Translating an idea of Hessenberg[12] into our notation, we start by defining a function of two variables X and Y, each variable ranging over the set of all points of L, excluding C, the values of the function ranging over the same domain. The function depends on four parameters, A, C, P, Q, such that P and Q are two distinct points outside of L, collinear with B. It is defined, as follows:

$$f_{ACPQ}(X,Y) = L \cdot [(X+P)\cdot(C+Q) + (Y+Q)\cdot(A+P)].$$

The theorem of Desargues implies that the function so defined does not depend upon the choice of the parameters P and Q. If R and S are two distinct points outside of L and collinear with B, then we have

$$f_{ACPQ}(X,Y) = f_{ACRS}(X,Y) \text{ for each } X \text{ and } Y.$$

Consequently, we shall write

$$f_{ABC}(X,Y) \text{ instead of } f_{ACPQ}(X,Y).$$

Each function so defined is seen to be what may be called associative; i.e., we have

$$f_{ABC}(X, f_{ABC}(Y, Z)) = f_{ABC}(f_{ABC}(X,Y), Z).$$

The function $f_{ABC}(X,Y)$ has B as indifferent element, i.e.,

$$f_{ABC}(B, X) = f_{ABC}(X, B) = X \text{ for each } X.$$

For each element X there exists an inverse element X', i.e., an element such that

$$f_{ABC}(X', X) = f_{ABC}(X, X') = B$$

If, in particular, $A = C$, then the function f_{CBC} is seen to be symmetric, i.e.,

$$f_{CBC}(X, Y) = f_{CBC}(Y, X) \text{ for each } X \text{ and } Y.$$

Finally, the functions f_{CAC} and f_{ABC} satisfy what may be called distributive laws. We have

$$f_{ABC}(X, f_{CAC}(Y, Z)) = f_{CAC}(f_{ABC}(X, Y), f_{ABC}(X, Z))$$

and

$$f_{ABC}(f_{CBC}(Y, Z), X) = f_{CBC}(f_{ABC}(Y, X), f_{ABC}(Z, X)).$$

These remarks show that for any choice of three mutually distinct points A, B, C, the points of the line L except C form a semi-field in the sense of the algebra of numbers with f_{CAC} and f_{ABC} as the field-operations, field-addition and field-multiplication, respectively, A and B being their indifferent elements. It is customary to write 0, 1, ∞ instead of A, B, C and to call $f_{\infty, 0, \infty}(X, Y)$ the algebraic sum of X and Y, $f_{0, 1, \infty}(X, Y)$ the algebraic product of X and Y.

The semi-field is a field, that is to say, the field multiplication satisfies the commutative law

$$f_{ABC}(X, Y) = f_{ABC}(Y, X) \quad \text{for each } X, Y$$

if and only if the law of Pappus, or what is equivalent, the so-called fundamental law of projective geometry holds in the plane. Since the law of Pappus implies the law of Desargues it is a sufficient basis to permit us what is usually called the "algebraization" of the projective plane, that is, the proof that the points on a line in our algebra of geometry behave like the numbers of a field in the algebra of numbers. In the usual way, we can introduce homogeneous coordinates, i.e., associate an ordered triple of elements of a field with each point, in such a way that the straight lines are given by linear equations.

In concluding this remark we should like to mention that instead of $f_{ABC}(X, Y)$ another function can serve us as a basis of the introduction of coordinates on the line L. Let

H, J, K be three lines of our plane, all three $\neq L$ and $H \neq J \neq K$ whereas $H = K$ is not excluded. Then we set

$$g_{HJK}(X, Y) = L \cdot [(X + H \cdot J) \cdot K + (Y + K \cdot J) \cdot H].$$

From Desargues' law it follows that $g_{HJK}(X, Y) = g_{EFG}(X, Y)$ for each X, Y if $H \cdot L = E \cdot L$, $J \cdot L = F \cdot L$, $K \cdot L = G \cdot L$. The operations g_{KHK} and g_{HJK} behave like addition and multiplication in a semi-field, $H \cdot L$ and $J \cdot L$ being their indifferent elements.

As one sees, the notation of algebra of geometry permits a precise and elegant expression of the most important functions and operations of projective geometry.

5. ALT'S THEORY OF AFFINE GEOMETRY

In our joint paper, F. Alt[13] developed a postulational theory of affine geometry coordinating the latter with projective geometry according to the general program outlined in the introductory sections of this lecture. If we postulate two operations, denoted by + and $\cdot$, which are associative, commutative, admit indifferent elements, are absorbent, i.e., satisfy the law

$$A + (A \cdot B) = A \cdot (A + B),$$

and satisfy the intercalation law mentioned in Section 3, then we cover ground which is common to projective and affine geometry.

It is only the dual intercalation law of projective geometry which fails in affine geometry. For let U be the affine plane. In this case, each hyperplane is a straight line. Let A be a line, H a line parallel to A, and B a point on A. Then we have $A \supseteq B \supseteq A \cdot H (= V)$, and yet $B \neq A$ and $B \neq A \cdot H$. However, we readily see that it is only one small part of the dual intercalation law which breaks down in affine geometry. The following greater part of the dual intercalation law holds both in projective and affine geometry: If H is a hyperplane, and A and B are any two elements such that $A \supseteq B \supseteq A \cdot H \neq V$, then $B = A$, or $B = A \cdot H$.

This reduced dual intercalation law and the other assumptions mentioned above which hold both in projective and affine geometry, together form a system of postulates from which the theory of dimensions common to the two geometries can be developed. In particular, the fundamental equality for dimension can be replaced by

$$\dim A + \dim B \geq \dim(A+B) + \dim(A \cdot B)$$

where the $=$-sign holds if $A \cdot B \neq V$.

From this common basis both projective and affine geometry branch off, either one being obtainable after the adjunction of a specific assumption. The projective assumption is: If H is a hyperplane and $A \supseteq B \supseteq A \cdot H = V$, then $B = A$, or $B = V$. The affine assumption is: If P, Q, R are points, no one of which is part of the sum of the two others, then there exists exactly one element L, such that

$$R \subset L \subset P + Q + R \text{ and } L \cdot (P+Q) = V.$$

The projective assumption can be expressed by saying that each element of a dimension >0 has a non-vacuous intersection with each hyperplane. Calling parallel to the hyperplane H each non-vacuous element whose intersection with H is $= V$, we can formulate the projective assumption also by saying: There are no elements parallel to hyperplanes except points.

The adjunction of the projective assumption makes the whole system of postulates self-dual, i.e., all assumptions remain true if $+$ and $\cdot$, and V and U are interchanged, provided, of course, that also the concepts derived from these operations are properly dualized; in particular, that $\subset$ and $\supset$ are interchanged, and points and hyperplanes change their roles. The affine assumption obviously is a rephrasing of Euclid's parallel postulate: There is exactly one line (L) parallel to a given line ($P+Q$), passing through a given point (R) not on the given line, and contained in the plane ($P+Q+R$) determined by the given line and the given point. This assumption essentially deals with lines in a plane. One would expect that in spaces of a dimension >2 it might be replaced by a postulate dealing with hyperplanes: There is exactly one hyperplane parallel to a given hyperplane and passing through a given point outside the given hyperplane. In fact, the last statement is a consequence of the affine assumption. But Alt showed by an example that it together with the assumptions common to affine and projective geometry does not imply the affine assumption.

The adjunction of the affine assumption leaves the system of postulates undual. This lack of duality is inherent to the affine space. The dual of the affine assumption (i.e., of Euclid's parallel postulate) does not hold.[14]

In this way, affine and projective geometry become two coordinated theories dealing exclusively with the undefined concepts of "flats," "joining," "intersecting."

As was mentioned at the end of section 1, one of the classical developments of the affine plane started with the projective plane together with an arbitrary line L in it, called the line at infinity, and then redefined the concepts of point and line by restricting them to points not on L, and to lines $\neq L$. Moreover, two lines are called parallel if they intersect at a point of L, in other words, if they do not intersect at any point in the new sense of this term. It should be clear that this development of affine geometry is, strictly speaking, not one in terms of the concepts of "flats," "joining," "intersecting" alone as is the algebra of affine geometry outlined in this section, but requires an additional undefined constant term, viz., that of one constant line.

6. THE CONCEPT OF BETWEENNESS IN AFFINE GEOMETRY

In our joint paper Alt and I only developed theorems in affine geometry dealing with joining, intersecting, and parallelism. Now one of the most fundamental concepts of affine geometry is betweenness. In terms of this relation and the concept of point the whole affine geometry can be developed, as Veblen[3] showed. What is the role of this relation in the algebra of affine geometry?

On a line L in the projective plane, two pairs of points X, Y and A, B are said to be conjugate if there exist two distinct points P, Q outside of L and collinear with A such that

$$L\cdot[(X+P)\cdot(B+Q)+(Y+Q)\cdot(B+P)]=A$$

or, what is equivalent in presence of Desargues' law, if for each pair of distinct points P, Q outside of L and collinear with A, the above formula holds. Another equivalent condition is that there exist two distinct points R, Q outside of L and collinear with B such that

$$L\cdot[(X+R)\cdot(A+Q)+(Y+Q)\cdot(A+R)]=B.$$

In the notation of the previous section these conditions can be expressed by saying that

$$f_{BAB}(X, Y)=A \text{ and } f_{ABA}(X, Y)=B, \text{ respectively.}$$

On the basis of the concept of conjugate pairs of points, Pieri[15] gave the following definition: Two pairs of points of

L, the pair A, B and the pair C, D, do not separate each other if on L there exists a pair of points X, Y which is conjugate both to A, B and to C, D.

On a straight line of the affine plane this definition leads to the following concept of betweenness: The point D lies between the points A and B if and only if the pairs A, B and C, D do not separate each other where C is the midpoint of A and B. This latter is defined by the condition that the pairs C, ∞ and A, B are conjugate where for a line M we write $L \cdot M = \infty$ if and only if L and M are parallel.

If we consider the ordinary plane of analytic geometry with real numbers as coordinates, then on each line Pieri's concept of betweenness has the same content as the ordinary concept. To illustrate this equivalence by a typical example, we consider the horizontal axis and the points with the abscissae -1 and $+1$. According to the ordinary definition a point lies between these two points if its abscissa lies between -1 and $+1$. According to Pieri's definition the point A with the abscissa a lies between -1 and 1 if and only if there exists a pair of points X, Y conjugate both to -1, 1 and 0, A since the origin 0 with the abscissa 0 is the midpoint of -1 and 1. In view of the fact that two pairs of points on the horizontal axis are conjugate if and only if the cross ratio of their abscissae is $= -1$, it is easily seen that if X has the abscissa x, then the pairs X, Y and -1, 1 are conjugate if and only if Y has the abscissa $1/x$. Thus the point A lies between -1 and $+1$ if there exists a real number x such that the pairs of numbers 0, a and x, $1/x$ have the cross ratio -1. As one readily verifies, a number x satisfies this condition if and only if

$$x^2 - \frac{2}{a}x + 1 = 0.$$

This equation has real roots if and only if a lies between -1 and $+1$.

While in the special case of the real straight line, Pieri's and the ordinary definition of betweenness are equivalent, the postulates of the general affine geometry are not sufficient to prove that Pieri's concept of betweenness has the most important properties of this concept, e.g., the properties that for three mutually distinct points of a line, at most one lies between the two other ones, and at least one lies between the two other ones. That for Pieri's betweenness neither of these two properties is a consequence of the assumptions of affine geometry is clear from the following two examples.

If we consider analytic geometry of the plane in which the coordinates of each point are complex numbers, then the postulates of affine geometry hold; and yet according to Pieri's definition for any three mutually distinct points of a line, each one lies between the two other ones. For instance, on the horizontal axis of this plane a point A with the abscissa a lies between the points -1 and $+1$ if and only if there exists a complex number x such that the quadratic equation

$$x^2-\frac{2}{a}x+1=0$$

is satisfied. But a complex root of this equation exists for each complex number a. Thus according to Pieri's definition, on the horizontal axis of the complex affine plane, each point of the horizontal axis lies between the points with the coordinates -1 and $+1$. In the same way we could show that each point of the horizontal axis lies between the points with the abscissae 0 and 1, and the points with the abscissae -1 and 0. Thus of the points -1, 0, 1 each lies between the two other ones.

If we consider the analytic geometry of the plane in which the coordinates of each point are rational numbers, then according to Pieri's definition the point with the abscissa $\frac{1}{2}$ does not lie between the points -1 and $+1$. For the equation

$$x^2 - \frac{2}{a}x + 1 = 0$$

would have to have a rational root for $a = \frac{1}{2}$ which it does not have. In a similar way one can show that of the points $-1, \frac{1}{2}, +1$, none lies between the two other ones.

While Pieri can define betweenness in terms of joining and intersecting, he has to make special assumptions in order to guarantee that the betweenness so defined has the properties which we require of a ternary relation in order to call it betweenness, e.g., that it satisfies the postulates about betweenness formulated by Veblen[3] in his development of affine geometry in terms of the undefined concepts of point and betweenness. Expressed in terms of joining and intersecting, these assumptions are rather complicated. The same holds for the concept of separation of pairs of points of a line in projective geometry. It can be defined in terms of the projective operations but only complicated assumptions concerning the latter ones would guarantee that the concept so defined has the usual properties.

7. THE CONCEPTS OF PERPENDICULARITY AND CONGRUENCY IN EUCLIDEAN GEOMETRY

Can the whole of Euclidean geometry including the theory of perpendicularity and congruency of segments and of angles be developed in terms of joining and intersecting?

One of the oldest and best foundations of Euclidean geometry is that due to Pieri in terms of only two undefined concepts, "point" and "two points are equally distant from a third point." Hilbert[1] took congruency of segments and angles among his many undefined concepts, and devoted one of his five groups of axioms to these concepts. Veblen[16] gave an elegant development in terms of "point," "between," and "congruent" (the latter applied to pairs of points). In addition, foundations have been given based on the assumption of the existence of a group of transformations with certain properties. Such foundations will not be discussed in this paper.

More closely related to the algebra of geometry is another paper in which Veblen[3] develops the whole of Euclidean geometry in terms of "point" and "between" alone. Following Pasch he first derives the theory of the affine plane, and then adjoins a line at infinity arriving in this way at a projective plane. Then he chooses an arbitrary elliptic involution on the line at infinity, i.e., a projective transformation on this line which leaves no point fixed, and for each point P transforms the image of P back into P. Finally he defines two lines of the affine plane to be perpendicular if they intersect the line at infinity at two points which correspond to each other by the chosen involution. Moreover, Veblen defines congruency of pairs of points and of angles with reference to the arbitrarily chosen involution.

The relation between this idea and the algebra of geometry is two-fold. We saw that the basic concepts of Veblen's theory, point and betweenness, can be defined in terms of joining and intersecting. Thus Veblen's theory would imply the possibility of founding the whole of Euclidean geometry on the two projective operations. Besides, the idea can be directly applied to a projective plane whose theory has been developed in terms of joining and intersecting. Instead of omitting a line and its points ("at infinity") as one theory of the affine plane does, we omit a line L on which an involution has been arbitrarily chosen. With reference to L and the involution on L we can define perpendicularity for the lines which are $\neq L$, and congruency of segments and angles in the same way as Veblen does on the basis of affine geometry. But are these really developments of Euclidean geometry in terms of joining and intersecting?

As we saw at the end of section 5, even the affine geometry developed in this way is, strictly speaking, based on more than the undefined concepts joining and intersecting, viz., on an undefined constant term L in addition to the projective operations. Still, with regard to the affine geometry we saw the possibility of a development which really is based on the concepts of joining and intersecting and on nothing else, viz., the algebra of affine geometry discussed in section 5.

For Euclidean geometry, however, not only the methods outlined above do not really provide us with foundations of the whole theory on the operations joining and intersecting, but, as Tarski recently[17] pointed out, there does not exist any system of postulates of Euclidean geometry in terms of point and betweenness (nor, as we may add, in terms of joining and intersecting) since congruency (and, we may add, perpendicularity) cannot possibly be defined in terms of point and betweenness (nor in terms of the two projective operations). This impossibility is a consequence of the following theorem due to Padoa:[18] If a system of propositions in terms of the undefined concepts $C, C_1, C_2, \cdots$ admits two interpretations (either one satisfying all the propositions) such that $C_1, C_2, \cdots$ have the same meaning in both interpretations, while C has different meanings in the two interpretations, then the concept C cannot be defined in terms of $C_1, C_2, \cdots$ on the basis of the system of propositions. (Reference to the basic propositions has to be made in the assertion of the non-existence of a definition in terms of $C_1, C_2, \cdots$ since it is these propositions which constitute the content of the otherwise undefined concepts $C_1, C_2, \cdots$.)

Now the plane Euclidean geometry developed in terms of flats, joining and intersecting, betweenness, parallelism, perpendicularity and congruency admits, e.g., the following two interpretations in the projective plane: One obtained from the projective plane by omitting a line L on which an elliptic involution I_1 has been chosen, and one obtained from the projective plane by omitting the same line L on which, however, a different elliptic involution I_2 has been chosen. While the concepts of flats, joining and intersecting, parallelism and betweenness have the same meaning in both interpretations, the concepts of perpendicularity and congruency have different meanings. For there are two lines perpendicular with regard to the involution I_1 which are not perpendicular with regard to I_2, and the situation with respect to congruency is similar. Hence, by virtue of Padoa's theorem, perpendicularity and congruency cannot be defined in terms of the concepts flat, joining and intersecting, parallelism and betweenness.

In concluding the remarks on Euclidean geometry I wish to point out why Tarski's remark has no analogue with regard to the affine geometry. It should not be believed that Padoa's theorem also yields the impossibility of defining parallelism and betweenness in terms of the projective operations. Indeed, the postulates of affine plane geometry have two different interpretations in the projective plane. E.g., one can be obtained by omitting from the projective plane a line L_1, the other by omitting from the projective plane a line L_2 ($\neq L_1$). Parallelism and betweenness have different meanings in the two interpretations. E.g., there are lines which are parallel in the first interpretation without being parallel in the second. Yet Padoa's theorem does not yield a proof of the impossibility of defining parallelism and betweenness in terms of the flats and the projective operations since also the concept of flat has different meanings in the two interpretations. Point means in the first interpretation "point of the projective plane outside of L_1," in the second interpretation "point outside of L_2."

As a matter of fact, it is possible to define parallelism and betweenness in terms of the projective operations on the basis of the postulates of affine geometry, as we have seen. What is impossible is a projective definition of the metric concepts on the basis of Euclidean geometry.

8. THE ALGEBRA OF BOLYAI-LOBACHEVSKY GEOMETRY

About two years ago the writer noticed[19] that it is possible to develop the whole of Bolyai-Lobachevsky geometry, including its metric parts dealing with perpendicularity and congruency, from assumptions about the operations of joining and intersecting only. In other words, instead of five groups of axioms, which, following Hilbert's development of Euclidean geometry, are frequently used as the basis of non-Euclidean geometry, one is sufficient. The contrast with the situation concerning Euclidean geometry outlined in the previous section seems remarkable.

Bolyai-Lobachevsky geometry is the theory of undefined concepts (points, lines, planes, etc., joining and intersecting, betweenness, parallelism, and congruency for segments and angles) satisfying all postulates about these concepts formulated in Euclidean geometry except the Euclidean parallel postulate, instead of which the following postulate holds: in a plane, to a line *L* through a point *P* not on *L*, there exist infinitely many lines not intersecting *L*; they form an angular sector in the pencil of lines through *P*; the two extreme lines of this angular sector likewise do not intersect *L* and are called the two parallels proper through *P* to *L*.

That this theory is consistent, if Euclidean geometry is, was proved around 1870 by Beltrami and Klein. Both of them found models for non-Euclidean geometry in the Euclidean space. Such a model is a system of entities, operations and relations, in the Euclidean space satisfying all the postulates made in non-Euclidean geometry about its undefined concepts. Later Poincaré and others devised further models for non-Euclidean geometry in the Euclidean space. For our purposes, illustrations by Klein's model will be most convenient.

In this model of the non-Euclidean plane the entities called points and lines are the points and open chords in the interior of an ellipse in the affine plane. The operations and relations in the model are defined as follows: The line "joining" two points is the open chord of the ellipse passing through the two points. The "intersection" of two lines is the point which the two chords have in common if they have any point in common; otherwise, they are called "non-intersecting" lines. If *P*, *Q*, *R* are three points on a line *L*, that is to say, points of an open chord of the ellipse, then we say that *Q* lies "between" *P* and *R* if it does so on the chord. If *P* and *Q* are two distinct points, and *O* and *R* are the two points of the ellipse which are the endpoints of the open chord passing through *P* and *Q*, then, using an idea due to Cayley, Klein calls "distance of *P* and *Q*" the absolute value of the logarithm of the cross ratio of the quadruple (*P*, *Q*; *O*, *R*). Two pairs of points

P, Q and P', Q' (on the same line or on different lines) are called "congruent" if the distances of P and Q, and of P' and Q' are equal. The operations and relations so defined satisfy all the postulates of Euclidean geometry, except Euclid's parallel axiom. To an open chord L of the ellipse through a point P not on the chord there are infinitely many chords not intersecting L. They form an angular sector in the pencil of all chords through P. The two extreme chords of this sector are the ones which have one endpoint in common with the chord L. Since this endpoint lies on the ellipse and not in its interior, these two extreme open chords do not intersect the open chord L and consequently play the role of the two parallels proper to L through P. Thus, Klein's model satisfies Bolyai-Lobachevsky's parallel postulate, and each pair of parallel lines in Klein's model is a pair of open chords of the ellipse which have one endpoint in common.

Now we are going to prove our statement made at the beginning of this section that the whole of non-Euclidean geometry can be based on assumptions about joining and intersecting.[19] For this purpose we introduce the following auxiliary concept: In a non-Euclidean plane three mutually non-intersecting lines L_1, L_2, L_3 will be said to form an asymptotic triangle if for each number $i = 1, 2, 3$, through each point P_i of L_i, there is exactly one line different from L_i which intersects neither of the other two lines. It is very easy to illustrate the concept of asymptotic triangle in Klein's model: an asymptotic triangle is a triangle whose vertices lie on the ellipse bounding the model. In terms of asymptotic triangles, we shall now characterize the relations of betweenness, parallelism, congruency, and convergence.

If P, Q, R are three points on a line, then Q lies between

P and R, if and only if there is an asymptotic triangle, two of whose sides pass through P and R, and such that each line through Q intersects at least one of the three sides of the triangle. The easiest way of proving that this condition is really necessary and sufficient for Q's lying between P and R, is again suggested by Klein's model. If Q does lie between P and R, then each line through Q will intersect at least one of the three sides of any triangle whose vertices lie on the ellipse and two of whose sides pass through P and R. If a point S is collinear with P and R, and does not lie between P and R, then there are infinitely many lines through S, none of which intersects any side of a given triangle whose vertices lie on the ellipse and two of whose sides pass through P and R.

Two lines L and M are parallel if and only if they do not intersect and there exists a third line N such that L, M, N form an asymptotic triangle. Again it is clear from Klein's model that this is a necessary and sufficient condition for parallelism. For in Klein's model two parallel lines are represented by open chords which have a common endpoint on the ellipse, and two such chords can be complemented by a third chord in such a way that the three chords form a triangle inscribed in the ellipse.

Let P, Q and P', Q' be two pairs of points on parallel lines, L and L' respectively. Let M be the line parallel to L through P' distinct from L', and let M' be the line through P parallel to L' distinct from L. Let P^* be the intersection of M and M'. In the same way, we call Q^* the intersection of the parallel to L' through Q and the parallel to L through Q'. In order that P, Q and P', Q' be congruent, it is necessary and sufficient that the line L^*, joining P^* and Q^*, is parallel to L and L'.

From Klein's model, the necessity and sufficiency of this condition is clear. For denote by *O* and *R* the two endpoints of the chord *L*, by *O′* and *R* the endpoints of the chord *L′*. Call *K* the line together with which the parallels *L* and *L′* form an asymptotic triangle. If *L** is parallel to *L* and *L′*, that is to say, passes through *R*, let *O** be the intersection of *K* with *L**. Then the four points *O**, *P**, *Q**, *R* are perspective with the points *O*, *P*, *Q*, *R* from the center *O′*, and with the points *O′*, *P′*, *Q′*, *R* from the center *O*. The cross ratio of *O**, *P**, *Q**, *R* is thus the same as the cross ratio of *O*, *P*, *Q*, *R* and *O′*, *P′*, *Q′*, *R*. Hence *O*, *P*, *Q*, *R* and *O′*, *P′*, *Q′*, *R* have the same cross ratio, which means that *P*, *Q* and *P′*, *Q′* are congruent. If, however, *L** is not parallel to *L* and *L′*, that is to say, if the chord *L** does not end in *R*, then the cross ratios *O*, *P*, *Q*, *R* and *O′*, *P′*, *Q′*, *R* are easily seen to be neither equal nor reciprocal and, hence, *P*, *Q* and *P′*, *Q′* are not congruent in this case.

If *P*, *Q* and *P′*, *Q′* lie on non-parallel lines *L* and *L′*, then the two pairs are called congruent if there exists a pair *P**, *Q** on a line *L** parallel to both *L* and *L′* which is congruent to both *P*, *Q* and *P′*, *Q′*. (For any two lines *L* and *L′* there exists a line *L** which is parallel to both *L* and *L′*.)

If *P*, *Q*, *R* are three non-collinear points, then a point *O* lies in the interior of the triangle *PQR* if and only if each line through *P* contains at least one point between *P* and *Q*, or between *Q* and *R*, or between *R* and *P*. A sequence of points $O_1, O_2, \cdots$ converges to the point *O* if and only if each triangle *PQR*, in whose interior *O* lies, contains almost all the points of the sequence.

Thus in non-Euclidean geometry we have given characterizations of all relations in terms of joining and intersecting. In order to develop the whole non-Euclidean geometry in these terms we proceed as follows: we use the statements characterizing betweenness, parallelism and congruency as definitions of these relations. For instance, we define two non-intersecting lines to be parallel if there exists a third line such that the three form an asymptotic triangle. We say that the point *Q* lies between the points *P* and *R* with which it is collinear, if for each asymptotic triangle two of whose sides pass through *P* and *R*, each line through *Q* intersects at least one of the three sides of the triangle. And so on.

Now it is clear that we can develop the whole of non-Euclidean geometry from postulates concerning the projective operations of joining and intersecting. At the worst, we might take the five groups of postulates from which traditionally non-Euclidean geometry is developed, and in each postulate of one of the last four groups replace each of the words "between," "parallel," "congruent," and "converges" by our definition of these relations in terms of joining and intersecting. In this way, we should get a system of postulates in terms of joining and intersecting only, from which the whole of non-Euclidean geometry could be developed.

But, of course, a system of propositions obtained in this way would be very cumbersome. Many of the postulates would be quite unnatural since, in their ordinary form, they are adapted to the relations of betweenness, congruency, etc., rather than to projective ideas. Thus the problem arose of formulating simple direct postulates about the operations of joining and intersecting from which the whole of non-Euclidean geometry can be developed. In what follows I shall outline the work of two of my students in this direction.

9. THE WORK OF JENKS

F. P. Jenks[20] in his Notre Dame thesis (1940) succeeded in deriving the theories of order and of parallelism in the non-Euclidean plane from eight simple postulates. The postulates are formulated in terms of the undefined concepts "point," "line," and "lies on," which can be derived from the operations of joining and intersecting in the usual way. Two lines are called intersecting if they are distinct and there exists a point P which lies on both.

I. If P and Q are any two distinct points, there exists exactly one line L such that both P and Q lie on L.

II. Each line contains at least five distinct points.

III. There exist at least three non-collinear points, i.e., points which do not lie on one line.

IV. If L, M are two distinct non-intersecting lines, and P is a point not on L or M, then there exists at least one line through P which intersects neither L nor M.

V. If L, M and N are any three mutually non-intersecting lines, and if there exists a line which meets L and M but not N, and another line which meets M and N but not L, then through each point of L there exists a line which meets M but not N.

VI. If P, Q, R are three distinct collinear points and if there exist two lines L and L' through P, two lines M and M' through Q, and two lines N and N' through R, such that

(1) M and N intersect, but neither intersect L, and

(2) L' and M' intersect, but neither intersect N',

then each line through Q intersects at least one line of every pair of intersecting lines which pass through P and R respectively.

VII. If L and M are any two distinct intersecting lines and P is any point not on L or M, then there exists at least one line through P which intersects L but not M.

VIII. If L is a line, and P a point not on L, then there exist at least two distinct lines through P not intersecting L.

It should be noted that the postulates I–VI are common to non-Euclidean, affine, and projective geometry. Postulates IV, V, VI are vacuously satisfied in the projective plane, where any two lines intersect, postulates V, VI are vacuously satisfied in the affine plane where for a given line L and a point P not on L there is not more than one line through P not intersecting L.

Postulates I–VII are statements common to non-Euclidean and affine geometry but Postulate VII does not hold in projective geometry.

Postulate VIII is the only one which fails both in projective and in affine geometry, and thus is the sole typically non-Euclidean assumption among the eight. It is hardly necessary to remark that it is a part of the Bolyai-Lobachevsky parallel postulate.

Jenks bases his theory of order on the following definition: Q lies between P and R (which is expressed by the symbol PQR) if P, Q, R are three distinct points and if every line through Q intersects at least one line of each pair of intersecting lines through P and R. This definition of betweenness differs from the one mentioned in the previous section in two respects. The first is that it does not assume that the three points P, Q, R are collinear. Jenks proves from the other assumptions that if Q is between P and R in his sense, then P, Q, R are by necessity collinear. The second difference which constitutes a definite improvement compared with the definition mentioned in the last section consists in the directness and finiteness of Jenks' definition. By this we mean that it is not based on the concept of asymptotic triangles and operates with intersecting lines, rather than with parallel lines.

From his eight mentioned postulates Jenks was able to derive the theories of order and of parallelism in the Bolyai-Lobachevsky plane.

The first result is that for the points of each line *L* the ternary relation *PQR* satisfies the six properties taken to be characteristic for a between relation. In development of the theory of order due to Pasch, these six properties were formulated by Huntington and Kline.[21] They read as follows:

1. If *PQR* (that is, *Q* lies between *P* and *R*), then *P*, *Q*, and *R* are distinct.
2. If *PQR*, then *RQP*.
3. If *PQR*, then not *QRP*.
4. If *P*, *Q*, *R* are any three distinct points of *L*, then either *PQR* or *RQP* or *QRP* or *PRQ* or *RPQ* or *QPR*.
5. If *PQR* and *PRS*, then *PQS* and *QRS*.
6. If *PQR* and *QRS*, then *PQS* and *PRS*.

If *L* contains at least five points, then Property 6 is a consequence of the other five properties, as E. H. Moore showed. Now Jenks' postulate II states the existence of at least five points. Properties 1, 2 and 4 are immediate consequences of the definition of betweenness and of postulate VI.

As a typical example of a proof in this theory we mention the way Jenks derives property 3. The proof is essentially based on the part of the non-Euclidean parallel postulate expressed in Jenks' assumption VIII. The property 3 implies, in particular, that of three distinct points at most one lies between the other two. That this property of the between-relation follows from a part of the non-Euclidean parallel assumption is remarkable when contrasted with the fact mentioned in Section 6 that the ordinary assumptions of affine geometry including Euclid's parallel postulate are not sufficient to prove this property for the between-relation of affine geometry as defined by Pieri.

Jenks' proof reads as follows: Let M be a line through Q distinct from the line L through P, Q, R. (The existence of such a line follows from Postulate III.) By virtue of VIII there exist two distinct lines N_1 and N_2 through R neither of which intersects M. From VII it follows that there exists a line K through P which intersects N_1 but not N_2. Since M does not intersect N_1, and Q lies between P and R, it follows that M must intersect K. Since N_2 is a line through R which intersects neither of the intersecting lines K and M, the point R does not lie between P and Q.

By similar ideas Jenks proves property 5. In addition to the six properties of linear order he derives the statement known as the axiom of Pasch and, furthermore, what may be called convexity and external convexity of the plane, i.e., the existence, for any two distinct points P and R, of a point Q such that PQR, and of a point S such that PRS. From these theorems the properties of the order of points in the plane (each line divides the plane into two parts, etc.) can be derived in the usual way.

Jenks bases his theory of parallelism on the following definition: Two non-intersecting lines L, M are parallel if there exists a point P between L and M such that through P there is at most one line which meets neither L nor M. Here the point P is said to lie between the non-intersecting lines L and M if there exist points Q and R on L and M respectively such that QPR. This definition is proved to be independent of P in the sense that each point lying between L and M has the same property as P. Obviously, the relation of parallelism is symmetric.

Let N be a line which intersects two parallels L, M in the points P, Q respectively, and R be a point of L distinct from P. We say that L and M are parallel on the side of N on which R lies if there exists a line through R intersecting N in a point between P and Q but not intersecting M. Here again the definition is shown to be independent of the point R. If L and M are parallel, and M and N are parallel, and K is a line intersecting each of the lines L, M, N, then L and N are said to be parallel to the line M on the same side of the transversal K if there exists a point P on M such that both L and M, and M and N are parallel on that side of K on which P lies. Using these concepts, Jenks proves the classical assumption that the relation of parallelism is transitive in a certain sense: If L and N are two lines which are parallel to M on the same side of a transversal, then L and N are parallel. Further, if L and M are two parallels to a line N and through a point P, then L and M are parallel to N on opposite sides of any line through P which intersects N, and hence there exist at most two parallels to any given line through a given point. This is half of the Bolyai-Lobachevsky parallel postulate. Independent of Jenks' eight assumptions is the other half, namely, that there exist at least two parallels to any given line through a given point. Independent of this other half is the assumption that any two non-parallel lines have four distinct common parallels. Assuming their existence we can prove from the other postulates that the four common parallels lie in different quadrants.

10. THE WORK OF ABBOTT

The law of Pappus does not play any role in Jenks' theory but is the basis of J. C. Abbott's deductions concerning the concept of congruency. Abbott proved that if the plane satisfies the law of Pappus, then the congruency of pairs of points, defined in terms of joining and intersecting in a manner similar to that of section 8, has the following properties: If P, Q is a pair of points on a line L and P' a point on a line L', then there exists a point Q' on either side of P' on L' such that P', Q' and P, Q are congruent. If for three points P, Q, R on a line L for which Q lies between P and R, and for three points P', Q', R' on a line L' for which Q' lies between P' and R', the pairs of points P, Q and P', Q' are congruent and the pairs Q, R and Q', R' are congruent, then the pairs P, R and P', R' are congruent.

In developing the point of view outlined in section 8, the writer tried to characterize the Bolyai-Lobachevsky geometry among the more general geometries studied by Minkowski which admit a plane model similar to Klein's model of the Bolyai-Lobachevsky plane but contained in the interior of any convex curve not necessarily an ellipse. It has been remarked[19] that the characterization can be formulated in terms of joining and intersecting, viz., by postulating Pascal's law for "asymptotic" hexagons which in Klein's model are represented by hexagons inscribed in the boundary curve of the model.

According to Abbott, it is in the theory of perpendicularity that we actually need Pascal's law in developing Bolyai-Lobachevsky geometry in terms of joining and intersecting. Abbott gives the following definition of perpendicularity in terms of the projective operations: Let L and L' be two lines intersecting at the point P. Let Q be an arbitrary point on L which is distinct from P. Let K and M be the two parallels to L' through Q. Let N be a parallel to either K or M through P, say, to K. Then we say that L and L' are perpendicular if L' passes through one of the two points of intersection of the two quadrangles of lines formed by the four common parallels to M and L and the four common parallels to N and L.

Now from Pascal's law it follows that if L' and L'' are both perpendicular to L and have a point in common, then $L' = L''$. This theorem implies as a corollary that if in the definition of perpendicularity we choose N as a parallel to M through P, then we get the same concept of perpendicularity; that is to say L and L' are perpendicular in the second sense if and only if they are perpendicular in the first.

If Pascal's law does not hold then, as Abbott shows by an example, all the other postulates may be satisfied and yet two distinct and intersecting lines may have a common perpendicular.

A further consequence of Pascal's law is the theorem that if L is perpendicular to L', then L' is perpendicular to L.

Next Abbott defines congruency of angles. An angle between the lines L, M intersecting at the point P is one of the four pairs of rays determined by P (one on L and one on M). Let α be an angle between L and M with the vertex P, and let α' be an angle between L' and M' with the vertex P'. Then let N be the perpendicular to L which at the same time is parallel to M and lies in the sector of the angle α. Similarly we define N'. Let Q be the intersection of N and L, and Q' the intersection of L' and N'. Then α and α' are said to be congruent if and only if the pairs of points P, Q and P', Q' are congruent.

By a repeated use of Pascal's law, Abbott proves that an angle between L and M is congruent to the angle between M and L with the same angular sector. Of his other results I mention only the theorem that any two right angles, i.e. angles between perpendicular lines, are congruent.

This research, which is still in progress, will result, I believe, in a system of fairly simple assumptions in terms of the projective operations for the whole Bolyai-Lobachevsky geometry.

NOTES

[1] *Grundlagen der Geometrie*, translated by E. J. Townsend, Chicago, Open Court Pub. Co.

[2] Veblen and Young, *Projective Geometry*, Boston, Ginn & Co.

[3] *Trans. Amer. Math. Soc.*, Vol. 5 (1904) p. 342.

[4] In *Memorie* and the *Atti Acc. R. Torino* (1895–1905).

[5] *Math. Ann.*, LXXIII (1913), p. 522.

[6] *Vorlesungen über neuere Geometrie*, 1882; new ed., Springer, 1922.

[7] *Jahresber. d. Deutsch. Math. Ver.*, XXXVII (1928), p. 309; *Ergebnisse e. math. Kolloquiums*, Vienna, 1 (1931), p. 28; *Annals of Math.*, Vol. 37 (1936), p. 456. This paper is written in collaboration with F. Alt and O. Schreiber. In the introduction the main points of view of this section were brought forth. See also the note, *Annals of Math.*, Vol. 38 (1937), p. 450, and the papers quoted in notes 10 and 11.

[8] *C. R. Congr. Intern. Math.*, Oslo, 1936, p. 297.

[9] *Annals of Math.*, Vol. 36 (1935), p. 743, and a series of publications on lattice theory.

[10] See the writer's paper "Axiomatique simplifiée de l'algèbre de la géométrie projective," *Comptes Rendus*, Paris, t. 206 (1938), p. 308. The weak distributive law was brought forth by O. Schreiber in his Vienna thesis incorporated in our joint paper in the *Annals of Math.*, Vol. 36 (1935).

[11] *Wien. Akad. Anz.*, 1928, No. 26; *Monatsh. f. Math. u. Phys.*, XXXVI (1929), p. 269; and *Ergebnisse e. math. Kolloquiums*, 1 (1931), p. 29.

[12] *Acta Math.*, XXIX (1905), p. 1.

[13] Section 5 of our joint paper in the *Annals of Math.*, Vol. 37, developing the algebra of affine geometry, is due to F. Alt (p. 481, footnote 13).

[14] We get a self-dual part of the affine plane if from the projective plane we omit a line and all points on it, and a point outside of this line and all the lines through it. Analytically each remaining point can be described by a pair of real numbers, x, y, not both $\neq 0$, each remaining line by an equation, $ax+by=1$.

[15] *Memorie Acc. R. Torino*, 1898, p. 24.

[16] "The Foundations of Geometry," in Monographs on Topics of Modern Mathematics, ed. by J. W. A. Young, New York, Longmans, Green & Co.

[17] Alfred Tarski, "Einige methodologische Untersuchungen über die Definierbarkeit der Begriffe," *Erkenntnis*, V (1935), p. 80.

[18] *C. R. 2e Congr. Intern. Math.*, 1902, p. 245, and *L'Enseignement Math.*, t. 5 (1908), p. 85.

[19] *Bull. Amer. Math. Soc.*, Vol. 44 (1938), p. 821; *Comptes Rendus*, Paris, t. 207 (1938), p. 458; *Proc. Nat. Acad. Sci.*, Vol. 24 (1938), p. 486.

[20] *Proc. Nat. Acad. Sci.*, Vol. 26 (1940), p. 277, and *Reports of Math. Colloquium*, Notre Dame, Ind., 2d ser., No. 1 (1939), p. 45, and Nos. 2 and 3 in press.

[21] *Trans. Amer. Math. Soc.*, Vol. 18 (1917), p. 301.

III

TOPOLOGY WITHOUT POINTS

1. THE FOUNDATIONS OF THE CONCEPT OF SPACE IN TOPOLOGY

THE three foundations of the concept of space in topology start with the concept of limit, the concept of neighborhood, and the concept of closure.

As a limit class Fréchet denotes a set of elements in which certain denumerable sequences of elements are distinguished. Each distinguished sequence is called convergent, and with each convergent sequence of elements exactly one element is associated, called the limit of the convergent sequence. It is assumed that each subsequence of a convergent sequence is convergent, and that each sequence all of whose elements are equal to one and the same element, is convergent and has this element as its limit. Instead of describing a limit class as a set in which certain sequences are distinguished, we can, in a more formal way, speak of a limit class if a set L and a subset C of d^L are given, where d^L denotes the set of all the denumerable sequences of elements of L, and the elements of C are called distinguished sequences. The two properties of the convergent sequences can be expressed as properties of the set C.

As a neighborhood class (or as he says, a topological space) Hausdorff denotes a set T (whose elements are called points) in which certain distinguished subsets are associated with the points. Each distinguished subset associated with a point x is called a neighborhood of the point x. It is assumed that

1. Each point x of T is element of at least one neighborhood of x, and each neighborhood of x contains the point x.
2. If U_x and V_x are two neighborhoods of x, then there exists a neighborhood of x which is a subset both of U_x and of V_x.
3. If y is a point contained in the neighborhood U_x, then there exists a neighborhood of y which is a subset of U_x.
4. For each pair of distinct points x and y, there exist two neighborhoods, U_x and U_y, which have no point in common.

Again it is possible to describe the distinction of certain subsets of T as neighborhoods by saying that besides the set T there is given a subset of the set of all subsets of T.

The space T is said to satisfy the second denumerability axiom if there exists a denumerable set of neighborhoods which is a basis for the open subsets of the space, i.e., such that each open set is sum of sets belonging to the basis.

As a closure class (or as he says a class $\bar{A}$) Kuratowski denotes a set in which with each subset A there is associated a set $\bar{A}$, called the closure of A. It is assumed that

1. The closure of the sum of two sets is equal to the sum of the closures of the two sets.
2. The closure of a set consisting of one point, consists of this one point.
3. The closure of the vacuous set is the vacuous set.
4. The closure of the closure of a set is equal to the closure of the set.

These three foundations of topology have in common that they are what may be called point set theoretical. By this we mean that each of them considers the space as a set of elements. Of course, it is a set with special properties distinguishing the space from an abstract set, viz., a set in which certain sequences of elements are distinguished, or a set in which subsets are associated with elements or with subsets in a certain way. But in all three cases it is assumed that the elements of the space, the points, are somehow given individually, and that the space character of the set consists in relations between, and properties of, certain sets of these elements. The same holds for other more or less related ways of introducing the concept of space in topology as well as for the fundamental concept of a metric space in which distance numbers are associated with the pairs of elements.

2. THE CLASSICAL FOUNDATIONS OF THE CONCEPT OF THE STRAIGHT LINE

In arithmetic, three ways of introducing the set of all real numbers, or, geometrically speaking, the straight line, on the basis of the theory of the rational numbers, have been known since the 70's of the last century. They are somewhat analogous to the three introductions of the concept of space in modern topology.

The one which is related to the concept of limit class is Cantor's definition of a real number as a convergent sequence (or, as he said, fundamental sequence) of rational numbers. A sequence of rational numbers $r_1, r_2, \cdots$ is convergent if for each natural number n there exists an index k_n such that $|r_i - r_j| < 1/n$, provided that $i \geq k_n$ and $j \geq k_n$. Two numbers $r_1, r_2, \cdots$ and $s_1, s_2, \cdots$ are said to be equal if $r_1, s_1, r_2, s_2, \cdots$ is a number. A $<$-relation and a limit concept can be introduced for the real numbers so defined. The straight line is the set of all these real numbers.

The way related to the concept of neighborhood class is the definition of a real number as a nested sequence of rational intervals whose lengths approach 0. By a rational interval, we mean a pair $I=(r_1, r_2)$ of rational numbers for which $r_1<r_2$. As the length of the interval I, we denote the rational number r_2-r_1. The interval $I=(r_1, r_2)$ is said to be contained in the interval $J=(s_1, s_2)$ if $s_1\leq r_1<r_2\leq s_2$. We say that I is completely contained in J if $s_1<r_1<r_2<s_2$. A sequence $I_1, I_2, \cdots$ of intervals is said to be strictly decreasing or nested if each interval I_k contains the following interval I_{k+1} completely. Two real numbers $I_1, I_2, \cdots$ and $J_1, J_2, \cdots$ are said to be equal if each I_k contains a J_l (and consequently almost all J_l completely), and each J_k contains an I_m (and consequently almost all I_m completely). Again an order relation and a limit concept can be introduced for the real numbers so defined. The set of all these numbers is called a straight line.

A way somewhat related to the concept of closure class is Dedekind's definition of a real number as a cut or upper section in the set of rational numbers. A set of rational numbers is called an upper section if together with each rational number, it contains all greater rational numbers. If for an upper section there exists a rational number (belonging or not belonging to the upper section) which is smaller than all the (other) numbers of the upper section, but greater than all the (other) numbers not belonging to the upper section, then we say that the upper section is a rational real number. If for an upper section there does not exist a rational number with these properties, then we say the upper section is an irrational real number. The set of all upper sections is called a straight line.

These three ways of introducing the straight line or the set of all real numbers do not presuppose the concept of a point or of a real number. On the basis of a denumerable set (the set of rational numbers or rational intervals) for whose elements certain relations are assumed to be given, (a <-relation for the rational numbers, a relation of containing for the rational intervals) they introduce the individual real numbers or points. The set of all of them is formed in order to enable us to talk, if necessary, about the straight line as a whole.

3. POINTS OF A GENERAL SPACE INTRODUCED AS SEQUENCES OF LUMPS

Huntington's foundation of Euclidean geometry is given in terms of spheres and a relation of inclusion.[1] Points are introduced as spheres which do not include any other spheres.

In his Paris thesis in 1923, Nicod[2] in continuation of ideas of Whitehead outlined a theory starting with "volumes," and a relation of "contained in the interior of." As an "abstractive class of volumes," he denotes a class of volumes such that, for any two volumes belonging to the class, one is contained in the interior of the other while no volume belonging to the class is contained in the interior of all the volumes of the class. If *A* and *B* are two such abstractive classes, then *A* is said to cover *B* if each volume belonging to the class *A* contains in its interior a volume belonging to the class *B*. As a point, Nicod defines an abstractive class *A* covered by every abstractive class which *A* covers. Like Huntington's elaborate theory, the rather sketchy developments of Nicod aim to lay a foundation of Euclidean geometry by introducing a definition of congruency for pairs of points.

In my book *Dimensionstheorie*[3] I pointed out the desirability of an introduction of the general concept of space in topology which is not point set theoretical in the sense of Section 1, but rather analogous to the introductions of the straight line in arithmetic, outlined in Section 2. I especially aimed to introduce the points of a space as nested sequences of what may be called pieces or lumps—analogous to the introduction of real numbers as nested sequences of rational intervals, and related to Hausdorff's concept of a neighboring class.

Such a topology of lumps seems to me to be closer to the physicist's concept of space than is the point set theoretical concept. For naturally all the physicist can measure and observe are pieces of space, and the individual points are merely given as the result of approximations.

Also, such a topology of lumps probably comes closer to the ideas of some philosophers than does the point set theoretical concept of space. It is true, in the philosophical literature the critique of the concept of the continuum as a set of points was most frequently directed against the introduction of the linear continuum, i.e., the straight line, although just in this case the introduction of points as nested sequences of rational intervals conforms with at least some of the philosophical exactions. But on somewhat better grounds some of the critical ideas might be directed against the traditional introductions of the general concept of space. To supply this latter concept with a foundation whose logical dignity equals that of the basis of the concept of the straight line, is the purpose of our theory.

We start with a partially ordered system of lumps which we shall denote by $U, V, \cdots$. They correspond to the rational intervals of arithmetic and to the neighborhoods of Hausdorff. We call the order relation defined for these lumps "completely contained in" and denote it by $\subset\subset$. The reason we speak of complete containing rather than containing, and use the symbol $\subset\subset$ rather than $\subset$ will be clear later, when we shall see that two lumps U and V

for which $U \subset\subset V$, correspond to two open sets in the topological space such that the closure of U is a subset of V—a relation described in topology by the symbol $U \subset\subset V$.[4]

About the partially ordered set of the lumps $U, V, \cdots$, we make the following assumption: Two elements U and V are identical if and only if for each element W the relations $W \subset\subset U$ and $W \subset\subset V$ are equivalent (that is to say, either both hold or both do not hold).

Now as points we define certain sequences of lumps $U_1, U_2, \cdots$, which are strictly decreasing; that is to say, for which $U_{k+1} \subset\subset U_k$ for each k. The points $U_1, U_2, \cdots$ and $V_1, V_2, \cdots$ will be called equal if and only if each U_i contains one V_j (and consequently almost all V_j) completely, and if each V_i contains one U_k (and consequently almost all U_k) completely. The point $U_1, U_2, \cdots$ is said to lie in the lump U if we have $U_i \subset\subset U$ for one (and thus for almost all) integers i.

It is clear that if we consider a strictly decreasing sequence $U_1, U_2, \cdots$ of open sets in a topological space, four cases can occur:

(1) *The sequence contracts to a point* p; that is to say, only the point p is contained in all open sets of the sequence, and each neighborhood of p (i.e., each open set containing p) contains a set U_i and consequently almost all U_i. E.g., this is the case for the strictly decreasing sequences of rational intervals of the straight line whose lengths converge toward 0, if by a rational interval we mean the set of all real numbers between two rational numbers.

(2) *The sets of the sequence have only one point in common without contracting to this point.* As an example, let us consider a space obtained from the straight line by omitting the point 1, and denote by U_n the open set consisting of the intervals $(-1/n, 1/n)$ and $(1-1/n, 1+1/n)$. The open

sets $U_1, U_2, \cdots$ have only the point 0 in common without contracting to the point 0. If we admit unbounded open sets, e.g., the ray consisting of all numbers $>n$ which we shall denote by (n, ∞), then even the ordinary straight line contains a decreasing sequence of open sets which have only one point in common without contracting to it, e.g., the sequence $U_1, U_2, \cdots$ where U_n denotes the sum of the interval $(-1/n, 1/n)$ and the ray (n, ∞).

(3) *The sets of the sequence have more than one point in common.* E.g., this is the case for a strictly decreasing sequence of rational intervals of the straight line whose diameters do not approach 0.

(4) *The sets of the sequence have no point in common.* E.g., this is the case if U_n denotes the set of all real numbers excluding 1 between $1-1/n$ and $1+1/n$ in the space obtained from a straight line by omitting the point 1, or if U_n denotes the ray (n, ∞) in the ordinary straight line.

In the case of the straight line, a number was associated with each rational interval, namely, its length, and the only condition that had to be imposed on a strictly decreasing sequence of rational intervals in order that it should define a real number was that the lengths of the intervals of the sequence should converge toward 0. This was necessary in order to exclude case (3). The cases (2) and (4) in the example of the rational intervals of the straight line are automatically excluded. More generally, it can easily be shown that in any compact space a sequence of open sets which have exactly one point in common, contracts to this point. Moreover, in a compact space the sets of a strictly decreasing sequence of open sets have at least one point in common. If the sequence $U_1, U_2, \cdots$ is decreasing without being strictly decreasing, then there need not exist a point common to all sets. E.g., even in the straight line the decreasing open intervals (0, 1), (0, 1/2), $\cdots$, (0, $1/n$), $\cdots$ have no point in common.

In my book I suggested the consideration of points of a space as certain decreasing sequences of lumps, without giving a criterion as to which sequences should be called points. I pointed out the desirability of formulating such a criterion in my colloquium in Vienna. In 1930 this problem was solved by A. Wald in a way that we shall outline presently.

At this point I should like only to mention that in order to obtain a space whose abstract properties are related to those of a straight line, we shall start with a system of lumps which has certain denumerability properties. On the other hand, it is clear that if we start with a set of lumps that is altogether denumerable (as I originally did), then we can not possibly get all topological types of spaces on the mere basis of the order relation for the lumps. For, the set of all partially ordered denumerable sets which are distinct (that is to say, no two of which are isomorphic) has the power of the continuum. Consequently, if we wish to describe the points of a space in terms of the order relation alone, we cannot get more than continuously many spaces —while the set of all types of topological spaces satisfying the second denumerability axiom, and even the set of all topological types of subsets of the straight line has a greater power than the continuum.

4. THE THEORY OF WALD

The main idea of Wald's theory[5] is a criterion which in a regular topological space satisfying Hausdorff's first denumerability postulate holds for a sequence in the case (1) and fails to hold for a sequence in one of the cases (2), (3), (4) mentioned in Section 3. In such a space a necessary and sufficient condition in order that the strictly decreasing sequence of non vacuous open sets U_1, U_2, $\cdots$ contract to a point is *that for each open set* V *which does not contain any of the* U_k *completely, and for each open set* W *which is completely contained in* V, *the set* W *is disjoint from almost all* U_k. Here two sets are called disjoint if they have no point in common. A space is called regular if for each point p lying in an open set U there exists an open set V containing p and such that $V \subset\subset U$. A space satisfies Hausdorff's first denumerability axiom if for each point p there exists a denumerable set of open sets containing p and such that each open set containing p contains one of the open sets of the denumerable set as a subset.

The necessity of the condition is easily verified. The interesting fact is its sufficiency. In order to prove that the condition is sufficient, we have to show that it does not hold in Cases (2), (3), and (4). Clearly, each sequence U_{i_1}, U_{i_2}, $\cdots$ extracted from a sequence satisfying the condition, satisfies the condition.

It is easily seen that the sets of a sequence U_1, U_2, $\cdots$ satisfying Wald's condition have at most one point in common. For, if the sets of a sequence U_1, U_2, $\cdots$ have two distinct points p and q in common, then we may choose an open set V containing p, but not q, and an open set W containing p and completely contained in V. Since q does not lie in V, the set V does not contain any of the sets U_k. Nevertheless, none of the sets U_k and W are disjoint since W contains p. Thus a sequence of open sets having more than one point in common does not satisfy Wald's condition. Hence, the condition excludes Case (3).

Next we show that the sets of a sequence U_1, U_2, $\cdots$ satisfying Wald's condition have at least one point in common. Let us assume that the sets of the sequence U_1, U_2, $\cdots$

have no point in common. Then, first of all, it is clear that if we pick out a point p_k of the set U_k, there does not exist any point p which is cluster of the sequence $p_1, p_2, \cdots$, that is to say, each neighborhood of which contains p_n for infinitely many n. For, on account of the strict decreasing of the sets $U_1, U_2, \cdots$ such a cluster point p would be common to all the sets U_k. Since the sequence $p_1, p_2, \cdots$ does not have any cluster point, it contains infinitely many distinct points. Thus, from $p_1, p_2, \cdots$, we may extract a sequence $q_1, q_2, \cdots$ of mutually distinct points. For each k, if q_k is the point p_{i_k}, we set $U_{i_k} = X_k$. Now we may choose open sets $V_1, V_3, V_5, \cdots$ such that of all the points $q_1, q_2, q_3, \cdots$ the set V_{2n-1} contains only the point q_{2n-1}. We may furthermore choose a neighborhood W_{2n-1} of q_{2n-1} which is completely contained both in V_{2n-1} and in X_{2n-1}. If now we set $V = V_1 + V_3 + V_5 + \cdots$ and $W = W_1 + W_3 + W_5 + \cdots$, then we see

(1) that V does not contain any of the sets X_k completely since it does not contain the points $q_2, q_4, q_6, \cdots$

(2) that W is not disjoint from any of the sets X_k since it contains the points $q_1, q_3, q_5, \cdots$

(3) that, as one can prove, the closure of W contains merely points of the closure of $W_1, W_3, W_5, \cdots$ and consequently is contained in V; that is to say, that W is completely contained in V.

Thus, the sequence $X_1, X_2, \cdots$ does not satisfy Wald's condition. From the assumption that the sets $U_1, U_2, \cdots$ have no point in common we derived that the sequence $U_1, U_2, \cdots$ contains a subsequence $X_1, X_2, \cdots$ not satisfying Wald's condition. Thus it does not satisfy the condition itself, and hence the condition excludes Case (4).

Finally, it is easily seen that Wald's condition also excludes Case (2).

In addition to the concept of complete containing, Wald's criterion refers also to the concept of two disjoint open sets. Two open sets are called disjoint if they have no point in common. Assuming that each open set does contain a point, we can define two open sets U and V to be disjoint by saying that there is no open set W which is completely contained in both.

On the basis of this criterion Wald solved the problem mentioned in Section 3 by formulating the following definition of a point on the basis of a partially ordered set of lumps: *A point is a strictly decreasing sequence of lumps,* $U_1, U_2, \cdots$ *such that for any lump* V *which does not completely contain any of the* U_k*, and any* W *completely contained in* V*, the lump* W *is disjoint from almost all* U_k. Here two lumps are called disjoint if there does not exist any lump completely contained in both.

As in Section 3, a point $U_1, U_2, \cdots$ is said to lie in the lump U if U contains completely a U_k (and consequently almost all U_k). Moreover, the point $U_1, U_2, \cdots$ is said to belong to the closure of the lump V if V is not disjoint from any of the U_k. By virtue of this definition it is clear that if W is completely contained in V, then each point of the closure of W is a point of V.

In order to prove that the set of all points defined as above form a regular topological space, it is necessary to assume that also conversely if W is not completely contained in V, there exists at least one point of the closure of W which is not a point of V. Under these assumptions one can prove that for each lump U the set of all points lying in U is an open set in the topological space obtained, and that for two lumps V and W, the closure of W is a subset of V if and only if the lump W is completely contained in the lump V in the sense of the partially ordered set. This is the reason we use the words "completely contained" and the symbol $\subset\!\subset$ rather than the word "contained" and the symbol $\subset$, to denote the order relation in the set of lumps.

We get a topological space satisfying Hausdorff's second denumerability axiom by assuming that the basic set of lumps consists of a denumerable system of lumps and lumps which in a symbolic way are sums of some of the lumps of the denumerable system. The points of the space can be defined as sequences of lumps belonging to the denumerable system; but in order to constitute points, a sequence must satisfy Wald's condition with regard to any two lumps V and W of the whole undenumerable set of lumps. (In proving that in a regular topological space Wald's criterion excludes Case (4) for a strictly decreasing sequence of open sets, we had to form sets V and W as sums of denumerably many open sets.) The reader interested in further details of Wald's theory is referred to his original paper.[5] The necessity of basing a theory which is to yield all regular topological spaces satisfying the second denumerability axiom on a non-denumerable set of lumps, follows from the remarks at the end of Section 3.

Here I should like to add to Wald's theory the following remark. In a compact space satisfying the denumerability axiom, that is to say, admitting a denumerable system of open sets which is equivalent to the system of all open sets, we can say: A strictly decreasing sequence $U_1, U_2, \cdots$ of open sets belonging to the denumerable system contracts to a point if and only if for each open set V *belonging to the denumerable system* which does not contain completely any of the U_k, and for each open set W *belonging to the denumerable system* which is completely contained in V, the set W is disjoint from almost all U_k. Moreover for each point of the space there exists a strictly decreasing sequence of open sets belonging to the denumerable system which satisfies the condition mentioned and contracts to the point. This remark shows that the theory of compact spaces can be developed on the basis of a denumerable partially ordered set of lumps. We simply have to define points as strictly decreasing sequences of lumps of the denumerable system, satisfying Wald's criterion with respect to the lumps of the denumerable system.

It seems certain to me that not only compact spaces, but also spaces of certain more general properties can be derived from a denumerable partially ordered set of lumps, and probably in a similar way. Of course, from the consideration at the end of Section 3 it follows that one cannot even hope to introduce all the spaces of the type in consideration unless their totality has a power not surpassing that of the continuum. It would be quite interesting to carry out this idea for Borelian and analytic spaces. Another question is what we can get by admitting as points transfinite sequences of lumps.

5. THE THEORY OF MOORE

Referring to previous lectures, R. L. Moore in 1935 published a foundation of the topology of the plane in terms of the concepts "piece" and "embedded in." We shall indicate the main points of contact and the main differences existing between the theory of Moore and the ideas outlined in the preceding sections, referring the reader to Moore's paper in Vol. 25 of the Fundamenta Mathematicae for the details of his development.

Moore starts with a partially ordered set of pieces. He assumes that for each integer n there exists a system of pieces with certain properties. A suggestive description of Moore's assumptions is obtained by calling "basic pieces of n^{th} degree of smallness" or briefly "pieces of n^{th} degree" the pieces which are members of the n^{th} system. By using this notation we do not mean to indicate that any metric assumptions about these pieces are explicitly made which would enable us to speak about a numerical upper bound of their size. But ultimately the pieces of the n^{th} system correspond to open sets whose diameters approach 0 with increasing n.

In this notation, Moore assumes:

(1) In each piece for each n there is embedded a basic piece of n^{th} degree of smallness.

(2) If the piece W is embedded in the piece V then there is an integer n such that each of the pieces of n^{th} degree which intersects W is embedded in V. (Moore says that two pieces U_1 and U_2 intersect if they are not disjoint in the sense of the theories outlined in the preceding sections; i.e., if there exists a piece embedded both in U_1 and U_2.)

(3) For each piece U and each n the set of the pieces of n^{th} degree embedded in U is coherent, i.e., cannot be split into two sets of pieces such that no piece of one of the sets intersects any of the pieces of the other set.

(4) For each piece U there exists a finite number of pieces of n^{th} degree, say $V_1 \cdots, V_k$, such that for a certain number m each piece of m^{th} degree intersecting U is embedded in one of the pieces $V_1 \cdots, V_k$.

As a point Moore defines *a decreasing sequence of pieces* $U_1, U_2, \cdots$ *such that for each* n, U_n *is one of the pieces of* n^{th} *degree.*

While the assumptions mentioned above represent merely a part of Moore's postulates, they are sufficient for a comparison of his ideas with the theories outlined in the preceding sections.

Driving toward a foundation of the topology of the plane, Moore makes rather special assumptions at the start. E.g., assumption (3) mentioned above essentially means that each piece is connected, thus making connected and locally connected the space derived from the set of pieces. Assumption (4) amounts to a Borel-covering theorem for the pieces thus making the closures of the pieces compact.

Clearly, Moore might have adopted the course of first drawing conclusions from the mere facts that the pieces form a partially ordered set having some general properties and gradually introducing more special assumptions. But even then his treatment would differ from the theory outlined in the preceding sections on account of his introduction of a topological degree of smallness, i.e., of the system of pieces of n^{th} degree for $n = 1, 2 \cdots$. It is these systems that enable him to exclude Case (3) mentioned in section 3, i.e., decreasing sequences of pieces whose members have more than one point in common. With their help he can define points simply as decreasing sequences of pieces of increasing degrees. We are reminded of the classical introduction of a real number as a nested sequence of rational intervals whose lengths approach 0. One of the main advantages of the theory outlined in the preceding section was a definition of points without the help of any (metric or topological) criterion of smallness—exclusively in terms of the relation "contained in."

Moreover, Moore defines as a point *each* nested sequence of pieces which get indefinitely small in his topological sense, and discards the possibility of Case (4), mentioned in section 3, viz., the possibility of a nested sequence of pieces whose elements have no point in common. Since this case can occur even in a regular topological space satisfying the second denumerability axiom, Moore in his definition implicitly assumes local compactness of the space to be derived, a procedure which again is justified by his aim, the topology of the plane. But just the exclusion of nested sequences without common point by an intrinsic criterion in terms of the relation "contained in" presented the main difficulty in introducing the general concept of space in the preceding sections.

Aside from these differences there are, of course, important similarities between the two approaches. In addition to the fundamental idea of describing the relation of "completely contained in" for open sets, and of defining points as strictly decreasing sequences, several details are common to the two theories. For instance, Moore's definition of a point pertaining to a piece *U* if one of the pieces defining the point is embedded in *U* (and consequently almost all of the pieces are), is the same as that of a point lying in a lump given in the preceding sections.

6. THE THEORY OF STONE

In the following sections we shall outline some recent theories which likewise avoid starting with given points, and instead introduce points as classes of given entities with given relations. But they move in directions different from the theories introducing points as nested sequences of lumps. Since they are better known than these latter theories our exposition will be brief.

M. H. Stone[6] starts with a Boolean ring or a distributive lattice with complementation, that is to say, with a system of entities for which two associative, commutative, totally linear, and distributive operations are defined, denoted by $+$ and $\cdot$, admitting indifferent elements V and U such that for each element A there exists a "complementary" element A' satisfying the conditions $A \cdot A' = V$ and $A + A' = U$.

Stone calls a "point" a set of elements, $\mathscr{A}$, which contains at least one element and does not contain all elements and has the following properties:

(S_1) *If* A *belongs to* $\mathscr{A}$, *and* $C \subset A$ (*i.e.*, $C+A=A$), *then* C *belongs to* $\mathscr{A}$.

(S_2) *If* A *and* B *belong to* $\mathscr{A}$, *then* $A+B$ *belongs to* $\mathscr{A}$.

(S_3) *If* $A \cdot B$ *belongs to* $\mathscr{A}$, *then at least one of the two elements* A, B *belongs to* $\mathscr{A}$.

A point so defined is said to be associated with the element A if the element A belongs to the set defining the point. The set of all points defined in this way is made a topological space in such a way that for each element of the Boolean ring, the corresponding point set, i.e., the set of all associated points, is an open subset of the space, and that the point sets associated with all the elements of the Boolean ring form a basis of the open subsets of the space. However, in the space so obtained the closure of a set consisting of exactly one point may contain other points. In order to exclude this possibility Stone has to make the following assumption about the Boolean ring:

(S_4) *No point* $\mathscr{A}$ *is proper subset of any other point* $\mathscr{B}$, the points considered as sets with the properties (S_1), (S_2), (S_3).

For two elements A and B of the Boolean ring the two associated point sets are complementary subsets of the space if and only if A and B are complementary elements of the Boolean ring. Since for each element the Boolean ring contains a complementary element whose corresponding point set is likewise open, it follows that each of the open sets of the basis of the topological space has a complementary set which is likewise open. From this we readily conclude that the topological space derived from the Boolean ring is totally disconnected. Moreover it can be shown to be bicompact.

Conversely, each given totally disconnected bicompact space can be obtained in this way.

Of the topological spaces homeomorphic with subsets of Euclidean spaces or the Hilbert space, those which are homeomorphic with a closed subset of Cantor's discontinuum and only those are included in Stone's theory.

7. THE THEORY OF WALLMAN

Instead of a Boolean ring, H. Wallman[7] starts with a mere distributive lattice without requiring that the lattice contain a complementary element of each of its elements. As a "point" Wallman defines each system of elements with the following properties:

(W_1) *Any finite number of elements of the system have a product which is* $\neq V$.

(W_2) *The system cannot be increased without losing property* (W_1), that is to say, for each element A which does not belong to the system there is a finite number of elements in the system whose product and A have the product V.

Comparing Stone's definition of a point with that of Wallman, we see that the latter is a special case of the former if we interchange the operations $+$ and $\cdot$, and, consequently, the relations $\subset$ and $\supset$. If we perform this change in Stone's definition it deals with a system $\mathscr{A}$ with the following properties:

(S_1') *If* A *belongs to* $\mathscr{A}$ *and* $C \supset A$, *then* C *belongs to* $\mathscr{A}$.

(S_2') *If* A *and* B *belong to* $\mathscr{A}$, *then* $A \cdot B$ *belongs to* $\mathscr{A}$.

(S_3') *If* $A+B$ *belongs to* $\mathscr{A}$, *then at least one of the two elements* A, B *belongs to* $\mathscr{A}$.

Now we readily see that a point in the sense of Wallman has these three properties, and in addition the property (S_4).

Wallman, like his predecessors, associates a point with the element A of the basic system (lattice) if the element

A belongs to the set of elements defining the point. He makes the set of all points a topological space in such a way that for each element of the lattice the corresponding point set, i.e., the set of all points associated with the element, is a closed subset of the space. Moreover, the point sets associated with all the elements of the lattice, form a basis for the closed subsets of the space, that is to say, each closed subset of the space can be represented as the product of closed sets of the basis.

However, the basis of the closed subsets of the space so obtained (with point set addition and multiplication) is not necessarily isomorphic with the lattice. A point set belonging to the basis may correspond to different elements of the lattice. In order to exclude this possibility, Wallman has to make the following assumption about the lattice:

(W_3) *For each pair of elements of the lattice,* A *and* B, *such that* A *not* $\subseteq B$, *there exists an element* C *of the lattice such that* $A \cdot C \neq V$ *and* $A \cdot C = V$.

The topological space so obtained can be shown to be bicompact. Conversely, each bicompact space can be obtained in this way.

8. THE THEORY OF MILGRAM

Stone starts with a Boolean ring, Wallman with a distributive lattice thus dispensing with complementation but retaining the two lattice operations. A. N. Milgram in a recent theory[8] starts with a mere partially ordered set satisfying some simple conditions thus dispensing with the lattice operations altogether.

A subset, S, of a partially ordered set P is called an upper section of P if for each element x belonging to S any element of P which is $>x$, belongs to S. Now let P be a partially ordered set containing a largest element u, for which there exists a system of upper sections, $\mathscr{S}$, with the following properties:

(M_1) *For any two elements* x *and* y *of* P *such that* x *not* $\leq y$, *the system* $\mathscr{S}$ *contains an upper section containing* x *but not containing* y.

(M_2) *No upper section belonging to* $\mathscr{S}$ *is a proper subset of any upper section belonging to* $\mathscr{S}$.

(M_3) *If* S *is an upper section belonging to* $\mathscr{S}$, *and* x *and* y *are two elements of* P *which are not contained in* S, *then* P *contains an element which is* $\geq x$ *and* $\geq y$ *and is not contained in* S.

Then Milgram calls "point" *each upper section of* P *belonging to* $\mathscr{S}$, and shows that T, the set of all these points, is a closure class whose closed sets have a basis isomorphic with P.

For each element x of P we form the set of those points (i.e., upper sections of P belonging to $\mathscr{S}$) to which the element x belongs. We shall denote this point set by $[x]$. For instance, $[u]$ is the set of all points, thus $[u]=T$. We further call $[P]$ the set of the point sets $[x]$ formed for all the elements x of P. By virtue of the condition (M_1) the set $[P]$, when partially ordered by the subset relation, is isomorphic with the partially ordered set P, that is to say, $[x]$ is a proper subset of $[y]$ if and only if $x<y$.

If now A is any point set, i.e., any subset of T, then Milgram calls "closure of A" the intersection of all point sets $[x]$ which contain A as a subset. There is at least one such set which does contain A as a subset, viz., $[u]=T$. If the space T contains more than one point, then the closure defined in this way satisfies Kuratowski's four postulates mentioned in section 1. The last one, i.e., that the closure of the closure of a set is equal to the closure of the set, is an immediate consequence of the definition of the closure.

The same is true for one half of Kuratowski's first postulate, viz., for the statement that the sum of the closures of two sets is a subset of the closure of their sum. The other half, viz., that the closure of the sum of two sets is a subset of the sum of their closures, is a consequence of the assumption (M_3). Also one half of Kuratowski's second postulate, viz., that a set consisting of exactly one point is a subset of its closure, immediately follows from the definition of closure. The other half, viz., that the closure of a set consisting of one point does not contain any other point, is a consequence of the assumption (M_2). Kuratowski's third postulate that the closure of the vacuous set is vacuous, means that the intersection of all point sets $[x]$ is vacuous. If the space contains more than one point, then this is a consequence of Kuratowski's second postulate and the definition of closure.

Each partially ordered set P satisfying the conditions (M_1), (M_2), (M_3) is thus isomorphic with a basis of the closed subsets of some closure class.

Conversely, let T be any given closure class, and P any basis of the set of all closed subsets of T, partially ordered by the subset relation. For each point p of T, let S_p be the set of all sets belonging to P which contain p. Obviously, for each point p the set S_p is an upper section of P. If we denote by $\mathscr{S}$ the set of the sections S_p formed for all points p of T, then the system $\mathscr{S}$ obviously satisfies the assumptions (M_1) and (M_2). It also satisfies the assumption (M_3). For if S_p is the upper section consisting of all the closed sets of the basis P which contain the point p, and if x and y are two closed sets of the basis P which do not belong to S_p, that is to say, do not contain p, then the set $x+y$ is a closed set not containing p although not necessarily belonging to P. But since P is a basis of the system of all closed subsets of T, the set $x+y$ is the intersection of closed sets belonging to P, and since $x+y$ does not contain p, not all sets of P containing $x+y$ contain p. Hence the basis P contains a set $\geq x$ and $\geq y$ and not containing p, thus not belonging to S_p.

It follows that (M_1), (M_2), (M_3) are both necessary and sufficient conditions for a partially ordered set to be isomorphic with a basis of the system of all closed subsets of some closure class.

This theory is entirely free of any reference to operations and deals with the order relations alone. If we look upon the theories of Stone and Wallman (the former dualized as in section 7), we see that they likewise introduce points as upper sections. If a Boolean ring satisfies Stone's condition (S_4), then his points are upper sections satisfying Milgram's conditions (M_1), (M_2), (M_3). The same is true for Wallman's points if the lattice satisfies condition (W_3).

By choosing a general partially ordered set as the basis of his theory Milgram not only provided the concepts of point and of space with a foundation free of irrelevant elements but obtained a means of deriving important topological theorems independent of the existence of points, that is to say, for entities more general than ordinary spaces. For in order to derive a space from a partially ordered set P he has to assume the existence of a system of upper sections of P, satisfying the three conditions (M_1) to (M_3) while important parts of topology go through without the last two of these conditions, provided that the system of upper sections satisfying (M_1) is denumerable or has a power not surpassing some preassigned cardinal number.

What is necessary in order to derive large parts of topology in a partially ordered set P is the existence of a system of upper sections of P containing for any two elements x and y

of P such that x not $\leq y$, an upper section to which x but not y belongs, and such that the power of the system lies under a preassigned cardinal number N. Such a system of upper sections Milgram calls a *strong separating system* (of a power $\leq N$) of the partially ordered set P. For many purposes it is even sufficient to know that there exists a system of upper sections of a power $\leq N$ which for any two elements x and y such that $x > y$, contains an element to which x but not y belongs. Such a system is simply called a *separating system* (of a power $\leq N$) of P. Clearly in these assumptions the accent lies on the bounded power of the separating systems. For the existence of a separating system in P whose power is equal to that of P is certain without any assumption. The set of upper sections of P formed for all the elements of P is an example of a strong separating system of P whose power equals that of P.

Let $U_1, U_2, \cdots$ be the elements of a separating system of P ordered in some way. If the separating system is denumerable, then we can order its elements in a sequence whose indices are natural numbers. If the separating system is non-denumerable, then the sequence $U_1, U_2, \cdots$ will necessarily be transfinite. Now let x_0 be a given element of P. If U_1 contains an element $> x_0$, then choose one of these elements and call it x_1. Otherwise, set $x_1 = x_0$. If U_2 contains an element $> x_1$, then choose one of them and call it x_2. Otherwise, set $x_2 = x_1$. Proceeding in this way, we get a sequence of elements $x_1, x_2, \cdots$ (which is denumerable if the separating system is denumerable, or transfinite otherwise) which has the following property: Any element y, which is $\geq x_n$ for each index n, is what we may call an upper bound above x_0; that is to say, it is $\geq x_0$ and P does not contain any element $> y$.

If P is what is called *inductive* of the power of the separating system, that is to say, if each increasing sequence $z_1, z_2, \cdots$ of elements of P whose power does not surpass that of the separating system, contains an element z such that $z \geq z_n$ for each element z_n of the sequence, then for each element x_0 there exists in P an upper bound above x_0.

If P is a set of closed (or open) subsets of a regular topological space satisfying the second denumerability axiom then there exists a denumerable separating system of P. If the $<$-relation means "is subset of" and if P is inductive of denumerable order, then by virtue of the last theorem each element of P is contained in an element of P which is not a proper subset of any element of P. Such an element is said to be saturated in P. Its existence constitutes Brouwer's so-called saturation theorem. Thus this theorem follows from Milgram's theorem (denumerable case).

Now let P be a set of subsets of a separable space and $<$ mean "contains as a subset." If P is inductive of denumerable order, and x_0 is an element of P, then x_0 contains an element of P, which does not contain any element of P as a proper subset. Such an element is said to be irreducible in P. Its existence constitutes Brouwer's so-called reduction theorem. If, for instance, P is the set of all sub-continua of a Euclidean space which contain two distinct points p and q, and $<$ means the subset relation, then p is inductive of denumerable order, by virtue of the theorem that the product of a sequence of continua containing p and q is a continuum containing p and q. The reduction theorem, in this case, yields the existence of an irreducible continuum joining p and q within each continuum joining p and q, a continuum being called irreducible between p and q if it has no proper subset which is a continuum containing p and q.

But the generalized saturation and reduction theorems are derived without the assumption of the existence of points.[9] In the same way Milgram derives the covering theorems of topology as a part of the theory of the partially ordered system of all subsets of a space without assuming that the space is a point set in the sense of the concepts discussed in Section 1. These are examples of a topology without points.

9. CONCLUDING REMARKS

The idea of dealing with the topological relations between the subsets or certain subsets of the space without considering them as point sets is of course related to the program which with regard to projective and affine geometry has been carried out by the algebras of these elementary geometries treated in the second lecture in this pamphlet. It is the same shift from the consideration of individual points or sets of points to the study of the mutual relations of spatial figures—in topology mainly of closed and open sets, in the elementary geometries of flats of different dimensions.

Another remark should clear up the relation between the theories discussed in sections 3-5, and those studied in sections 6-8. The former ones introduce points as nested sequences of lumps after the model of the introduction of real numbers as sequences of rational intervals—the latter ones introduce points as sets of lumps which if applied to the case of the straight line would yield a definition of a real number as the set of all open rational intervals containing the number.

A transition from the "sequence" definition to the "set" definition of point can be made by forming, for any sequence of lumps defining a point, the set of all lumps which contain at least one lump of the sequence. A transition from the "set" definition to the "sequence" definition can be made if the underlying partially ordered set is denumerable.

Compared with the earlier sequence theories, the method of Stone and Wallman has the advantage of being applicable to spaces in which the former theories (at least in their original form, dealing with denumerable sequences) would not work, and Milgram's theory even comprises entities which are not point spaces at all. Obviously, the sequence theories work only in spaces satisfying Hausdorff's first denumerability axiom. While this condition holds in the most important spaces, in particular, in all metrizable spaces, it is desirable for a foundation of topology to comprise still more general spaces. Of course, by admitting transfinite sequences of lumps we should get sequence theories which would also comprise more general spaces, but probably at the expense of simplicity. Moreover, from the algebraist's point of view, the set definition of points may seem preferable because of the analogy it establishes with the theory of ideals—in fact, Stone and Wallman use the terminology of the theory of numbers in formulating their definitions of points. However, Milgram's theory, as well as the sequence theories, seems to indicate that essentially the definition of point is based on the relation of partial order rather than on the lattice operations from which the analogy with the ideals is derived.

On the other hand, in those cases in which the sequence definition of points works, it operates with a minimum of logical machinery. It is true that it has to complement the definition of a point by a definition of the identity of two points, but the set definitions of points need equivalent completeness postulates (S_4), (W_2), and (M_2). From the point of view of reducing the set theoretical character of the foundation, the sequence definition of points is obviously preferable to any set definition.

At any rate, there have been various attempts in the topological literature essentially towards the same end, namely, that of eliminating points as basic concepts, and yet so far these attempts have been quite unrelated. For some time it has seemed desirable to me to coordinate these ideas. The present lecture is intended to carry out this synthesis.

In concluding, I should like to point out that even the introduction of points as nested sequences of lumps somehow transcends what can be observed in nature. For, by a lump, we mean something with a well defined boundary. But well defined boundaries are themselves results of limiting processes rather than objects of direct observation. Thus, instead of lumps, we might use at the start something still more vague—something perhaps which has various degrees of density or at least admits a gradual transition to its complement. Such a theory might be of use for wave mechanics.

KARL MENGER

NOTES

[1] *Math. Ann.*, LXXIII (1913), p. 522.

[2] *La géométrie dans le monde sensible*, Paris, Alcan, 1924. An English translation in J. Nicod, *Foundations of Geometry and Induction*, New York, Harcourt, Brace & Co., 1930.

[3] *Dimensionstheorie*, Leipzig, Teubner, 1928, p. 15.

[4] In my book I defined points as decreasing sequences of open sets, not as strictly decreasing sequences, as is stated in this lecture in anticipation of Wald's method.

[5] *Ergebnisse e. Math. Kolloquiums*, Vienna, 3, 1932, p. 6.

[6] *Trans. Amer. Math. Soc.*, Vol. 40 (1936), p. 37, and Vol. 41 (1937), p. 374.

[7] *Annals of Math.*, Vol. 39 (1937), p. 112.

[8] *Reports of Math. Colloquium*, Notre Dame, Ind., 2^{d} ser., No. 1 (1939), p. 18; No. 2 (1940), p. 1; *Proc. Nat. Acad. Sci.*, Vol. 26 (1940), p. 291.

[9] Milgram also gives a proof without the use of transfinite induction (although, of course, using the axiom of choice). While Kuratowski's method of elimination of transfinite numbers [*Fundamenta Math.*, III (1922), p. 76] does not yield Milgram's theorem about partially ordered sets, the latter contains Kuratowski's theorem as a special case. Milgram's theorem that a partially ordered set which is inductive for *all* powers, contains for each element x_0 an upper bound above x_0, contains also the maximum principle formulated for algebraic purposes by M. Zorn in his paper, "A Remark on Method in Transfinite Algebra," *Bull. Am. Math. Soc.*, Vol. 41 (1935), p. 667.

R.R. DONNELLEY & SONS COMPANY · THE LAKESIDE PRESS · CHICAGO

CONTENTS

Vol. XXVI January, 1939 No. 1

PAGE

CAVELIER DE LA SALLE, FOUNDER OF THE FRENCH EMPIRE IN AMERICA—Elaboration of an address delivered on the occasion of the visit of the French National Mission Cavelier de la Salle to the Rice Institute, April 2, 1937, by GABRIEL LOUIS-JARAY, Director of the Mission, and President of the Executive Commission of the Comité France-Amérique; translated by Marcel Moraud, Agrégé de l'Université de France, Docteur ès Lettres (Paris), Professor of French at the Rice Institute.

I. A Portrait of de la Salle 1

II. De la Salle's Vision of a French Empire in America 4

III. Earliest Explorations on the Great Lakes 11

IV. The Expedition of 1682 to the Gulf of Mexico 22

V. De la Salle's Last Expedition, 1684-1687. 29

VI. Tragedy and Death of de la Salle 37

Vol. XXVI April, 1939 No. 2

A DESIGN FOR LIVING—Rockwell Lectures on Religious Subjects, delivered at the Rice Institute, April 25, 26, and 27, 1939, by HARRIS ELLIOTT KIRK, D.D., LL.D., Minister of the Franklin Street Presbyterian Church of Baltimore, Maryland.

I. The New Element in the Modern Temper 43

II. The Divine Purpose 61

III. The Human Response 81

Vol. XXVI July, 1939 No. 3

WHAT IS OUR IDEA OF A UNIVERSITY?—A sermon preached to the twenty-fourth graduating class of the Rice Institute by MONSIGNOR OLIVIER MAURAULT, P.D., D.C.L., LL.D., Rector of the University of Montreal . . 109

THINKING FOR ONESELF—A stenographic report of an address delivered at the twenty-fourth commencement convocation of the Rice Institute by GEORGE E. VINCENT, Ph.D., LL.D., President of the Rockefeller Foundation, retired 124

Vol. XXVI October, 1939 No. 4

VITAMINS IN HUMAN NUTRITION—A series of current popular lectures by ASA CRAWFORD CHANDLER, Ph.D. (California), Professor of Biology at the Rice Institute.

I. Vitamins: Spark Plugs of the Body 143

II. Vitamin A 157

III. The "B" Family—Vitamin B_1, Mother of the "B" Family 169

IV. The Offspring of the "B" Family—Riboflavine, Nicotinic Acid, etc. . . . 185

V. Vitamin C 196

VI. Vitamin D, the Sunshine Vitamin, and Other Fat-Soluble Vitamins . . . 210